Naval Medical Research Institute
503 Robert Grant Avenue
Silver Spring, Maryland 20910-7500

NMRC 2004-002 September 2004

STATISTICAL VALIDATION OF MUTUAL INFORMATION CALCULATIONS: COMPARISONS OF ALTERNATIVE NUMERICAL ALGORITHMS

C.J. Cellucci
A.M. Albano
P.E. Rapp

Bureau of Medicine and Surgery
Department of the Navy
Washington, DC 20372-5120

Naval Medical Research Institute
503 Robert Grant Avenue
Silver Spring, Maryland 20910-7500

NMRC 2004-002 September 2004

STATISTICAL VALIDATION OF MUTUAL INFORMATION CALCULATIONS: COMPARISONS OF ALTERNATIVE NUMERICAL ALGORITHMS

C.J. Cellucci
A.M. Albano
P.E. Rapp

Bureau of Medicine and Surgery
Department of the Navy
Washington, DC 20372-5120

Approved for public release;
Distribution is unlimited

NOTICES

The opinions and assertions contained herein are the private ones of the writer and are not to be construed as official or reflecting the views of the naval service at large.

When U.S. Government drawings, specifications, and other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever. The fact that the Government may have formulated, furnished or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

Additional copies may be purchased from:

Office of the Under Secretary of Defense (Acquisition & Technology)
Defense Technical Information Center
8725 John J. Kingman Road, Suite 0944
Ft. Belvoir, VA 22060-6218

Federal Government agencies and their contractors registered with the Defense Technical Information Center should direct requests for copies of this report to:

TECHNICAL REVIEW AND APPROVAL
NMRC 2004-002

This technical report has been reviewed by the NMRC scientific and public affairs staff and is approved for publication. It is releasable to the National Technical Information Service where it will be available to the general public, including foreign nations.

RICHARD B. OBERST
CAPT, MSC, USN
Commanding Officer
Naval Medical Research Center

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-01-0188

1. REPORT DATE *(DD-MM-YYYY)*
September 2004

2. REPORT TYPE
Technical Report

3. DATES COVERED *(From - To)*
2002-2003

4. TITLE AND SUBTITLE
Statistical validation of mutual information calculations: Comparisons of alternative numerical algorithms

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
601135N

6. AUTHORS
C.J. Cellucci, A.M. Albano, and P.E. Rapp

5d. PROJECT NUMBER
4508

5e. TASK NUMBER
.518

5f. WORK UNIT NUMBER
A0247

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Medical Research Center
(Code 00)
503 Robert Grant Ave.
Silver Spring, Maryland 20910-7500

8. PERFORMING ORGANIZATION REPORT NUMBER
2004-002

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Bureau of Medicine and Surgery
(Med-02)
2300 E. Street, N.W.
Washington, DC 20372-5300

10. SPONSOR/MONITOR'S ACRONYM(S)
BUMED

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
DN241126

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Given two time series X and Y, their mutual information, I(X,Y)=I(Y,X), is the average number of bits of X that can be predicted by measuring Y and vice versa. In the analysis of observational data, calculation of mutual information occurs in three contexts, identification of nonlinear correlation, determination of an optimal sampling interval particularly when embedding data, and in the investigation of causal relationships with directed mutual information. In this report a minimum description length argument is used to determine the optimal number of elements to use when characterizing the distributions of X and Y. However, even when using partitions of the X and Y axis indicated by minimum description length, mutual information calculations performed with a uniform partition of the XY plane can give misleading results. This motivated the construction of an algorithm for calculating mutual information that uses an adaptive partition. This algorithm also incorporates an explicit test of the statistical independence of X and Y in a calculation that returns an assessment of the corresponding null hypothesis.

15. SUBJECT TERMS
nonlinear correlations, numerical algorithm, mutual information

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Diana Temple
					19B. TELEPHONE NUMBER *(Include area code)*
Unclass	Unclass	Unclass	Unclass	54	301.319.7642

TABLE OF CONTENTS

Summary

Given two time series X and Y, their mutual information, I(X,Y)=I(Y,X), is the average number of bits of X that can be predicted by measuring Y and vice versa. In the analysis of observational data, calculation of mutual information occurs in three contexts, identification of nonlinear correlation, determination of an optimal sampling interval particularly when embedding data, and in the investigation of causal relationships with directed mutual information.

In this report a minimum description length argument is used to determine the optimal number of elements to use when characterizing the distributions of X and Y. However, even when using partitions of the X and Y axis indicated by minimum description length, mutual information calculations performed with a uniform partition of the XY plane can give misleading results. This motivated the construction of an algorithm for calculating mutual information that uses an adaptive partition. This algorithm also incorporates an explicit test of the statistical independence of X and Y in a calculation that returns an assessment of the corresponding null hypothesis.

The previously published Fraser-Swinney algorithm for calculating mutual information is described. This algorithm includes a sophisticated procedure for local adaptive control of the partitioning process. When the Fraser and Swinney algorithm and the algorithm constructed here are compared, they give very similar numerical results. Detailed comparisons are possible when X and Y are correlated jointly Gaussian distributed because an analytic expression for I(X,Y) can be derived for that case. Based on these tests, three conclusions can be drawn. First, the algorithm constructed here has an advantage over the Fraser-Swinney algorithm in providing an explicit calculation of the probability of the null hypothesis that X and Y are independent. Second, the Fraser-Swinney algorithm is the more accurate of the two algorithms when large data sets are used. With smaller data sets, the Fraser-Swinney algorithm reports structures that disappear when more data are available. Third, the algorithm constructed here requires about 0.5% of the computation time required by the Fraser-Swinney algorithm.

I. Introduction

Given two time series $\{X\}=\{x_1,x_2,\cdots\cdots x_{N_D}\}$ and $\{Y\}=\{y_1,y_2,\cdots\cdots y_{N_D}\}$, their mutual information, I(X,Y), is the average number of bits of $\{X\}$ that can be predicted by measuring $\{Y\}$. It can be shown that this relationship is symmetrical, I(X,Y)=I(Y,X). A mathematical definition of mutual information and a demonstration of this property is given in the first appendix. This appendix includes a summary of the principal mathematical properties of I(X,Y). A more systematic presentation is given in Cover and Thomas (1991). In the analysis of observational data, calculation of mutual information occurs in three contexts, identification of nonlinear correlation, determination of an optimal sampling interval particularly when embedding time series data, and in the investigation of causal relationships with directed mutual information.

Mutual information can be used to identify and quantitatively characterize relationships between data sets that are not detected by commonly used linear measures of correlation. Figure 1 recapitulates an example shown in Mars and Lopes da Silva (1987) and displays three data set pairs. The first shows x_i when x_i=-3 to +3 in steps of 0.0006 plotted against ε_i, a random normally distributed variable with zero mean and unit variance. The second element of Figure 1 shows x_i versus $x_i+.2\varepsilon_i$ where ε_i is the previously used random variable. In the third example of Figure 1, $y_i=x_i^2+.2\varepsilon_i$. Four measures were calculated with ten thousand element data sets. The first was the linear correlation coefficient r (Press, et al., 1992). The probability of the null hypothesis of zero linear correlation also was calculated. A small value of P_{Null} indicates a high degree of linear correlation. The Spearman rank order correlation r_S and the probability of the corresponding null hypothesis of non-correlation was calculated. If P_{Null} is small and r_S is positive, a positive correlation has been detected. If P_{Null} is small and r_S is negative, anti-correlation has been detected. Kendall's tau, a nonparametric measure of correlation, and its associated P_{Null} were calculated. This set of calculations also incorporated estimation of mutual information between $\{X\}$ and $\{Y\}$ using an algorithm that will be described in a subsequent section. That section also includes a description of the procedure used to calculate the probability of the null hypothesis of statistical independence.

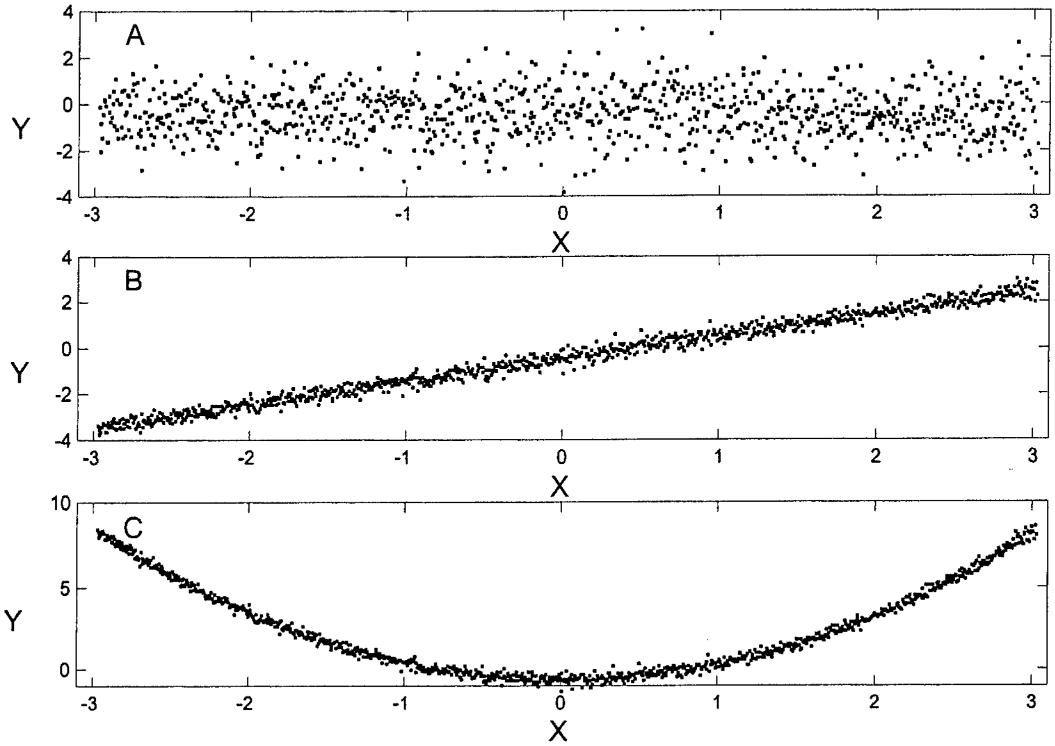

Figure 1. Data sets used in the correlation study of Table 1. In each case, x varies from –3 to +3 in steps of .0006. A. $y_i = \varepsilon_i$, a normally distributed random variable with zero mean and unit variance. B. $y_i = x_i + .2\varepsilon_i$ C. $y_i = x_i^2 + .2\varepsilon_i$

The results are shown in Table 1. In the case of normally distributed random numbers, all four measures behave in a manner that is consistent with our qualitative understanding of the word correlation. Measures r, r_S, τ, and I(X,Y) are small and the probability of the null hypothesis of zero correlation or, in the case of mutual information, statistical independence is high. Similarly, in the case of calculations with linearly correlated noise the results are consistent with expectations. The correlation measures are very nearly equal to one and the value of mutual information is high. The corresponding probability of the null hypothesis is numerically indistinguishable from zero in each case.

The results obtained in the case of parabolic correlation merit closer inspection. The first three measures r, r_S, and τ are small and the corresponding P_{Null} values are high, which indicates that no correlation was detected. In contrast, the value of mutual information is high, essentially equal to that obtained using linearly correlated data, and the probability of the null hypothesis of statistical independence is zero. Upon reflection it is seen that this is as it should be. Mutual information, I(X,Y), is the average number of bits of y that can be predicted by measuring x. Though the relationship between

{X} and {Y} in the third example is not linear, the relationship does confer a significant predictive capacity. This is reflected in the high value of I(X,Y) and in the low value of P_{Null}.

Table 1. Correlation Analysis

	Pearson r	Pearson P_{Null}	Spearman r_S	Spearman P_{Null}	Kendall's Tau	Kendall's P_{Null}	I(X,Y)	I(X,Y) P_{Null}
Normally Distributed Random	-.0037	.7112	-.0040	.6854	.0027	.6845	.1356	.7851
Linearly Correlated	.9934	0.	.9936	0.	.9270	0.	2.9186	0.
Parabolically Correlated	.0001	.9912	$<10^{-4}$	.9928	$<10^{-5}$	.9989	3.0304	0.

Mutual information is thus seen to be a nonlinear generalization of the concept of linear correlation. Beginning with the pioneering work of Callaway (Callaway and Harris, 1974) and Mars and Lopes da Silva (Mars, et al., 1985, 1987), mutual information has been used in studies of nonlinear correlations in multichannel EEGs. The investigations indicate that mutual information estimates can be used to discriminate between focal and generalized seizures. Additionally, in the case of focal seizures, the method can be used to identify the location of the epileptogenic focus (Mars, et al., 1985). Generalizations of the procedure in the form of directed mutual information will be considered presently.

Mutual information estimates also can be used to determine an appropriate sampling interval, T_S, which is the time between consecutive measurements of a time series. Many of the calculations presented here will be calculations directed to this question. The selection of an appropriate sampling interval is an important consideration when the quantitative methods of dynamical analysis are applied to time series data. On first consideration, one might suppose that the smallest possible T_S would be the best option. While this may be a reasonable approach during data acquisition, this strategy can fail during analysis because calculations with over-sampled data can produce misleading results (Rapp, et al., 1993). Historically, calculation of the autocorrelation time, the time required for the autocorrelation function to drop to 1/e of its initial value, has been used to establish an approximate sense of the time scale corresponding to significant changes in a time series' behavior. However, as we have seen in the preceding calculations, linear measures can give an incomplete characterization of behavior. This recognition has motivated the calculation of lagged mutual information.

Let {X} be the original time series, and let time series {Y} be the same time series shifted by time a lag, that is, $y_i = x_{i+\text{Lag}}$. The mutual information $I(X_i, X_{i+\text{Lag}})$ is then calculated. This process is repeated, and $I(X_i, X_{i+\text{Lag}})$ is determined as a function of Lag. In order to get the most new information from a measurement, we want to take the next measurement when there is maximum uncertainty in the relationship between {X} and {Y}. Recall that I(X,Y), which is symmetrical I(X,Y)=I(Y,X), is the average number of bits of Y that can be predicted by measuring X and vice versa. Therefore, the maximum uncertainty in the relationship between {X} and {Y} will occur at a minimum of $I(X_i, X_{i+\text{Lag}})$. This indicates that T_S should be set equal to a value of Lag that gives a minimum of $I(X_i, X_{i+\text{Lag}})$. It can be further argued, for example see Fraser and Swinney (1986), that among the many different minima of $I(X_i, X_{i+\text{Lag}})$, the sampling interval should correspond to the first minimum of $I(X_i, X_{i+\text{Lag}})$. This is particularly true when, as is often the case, chaotic systems are being investigated since the turbulent mixing of a chaotic system will cause an unacceptable loss of structure if T_S is too large.

Estimation of mutual information is often required when embedding dynamical data. In the simplest case, an analysis based on embedded data begins with a scalar time series {X}. The elements of {X} are then used to form an m-dimensional set $\{Z\} \in \Re^m$ with the construction

$$Z_j = (x_j, x_{j+\text{Lag}}, x_{j+2\text{Lag}}, \cdots\cdots x_{j+(m-1)\text{Lag}})$$

The analysis then continues with the investigation of the geometrical properties of {Z}. The motivation for this approach follows from the Takens-Mañé embedding theorem (Takens, 1980; Mañé, 1980), which shows that if the conditions of the theorem are met, then an intimate relationship exists between {Z} and the dynamical system which generated the observed times series {X}. This theorem requires the assumption that set {Z} is dense. This can never be satisfied with finite data sets. However, while recognizing the approximate nature of the analysis, an investigation of {Z} can in some instances provide significant information about the underlying generator. A crucial operation difficulty is encountered when embedding finite observational data sets. Embedding parameters m and Lag must be chosen. This choice is crucial to the success of the subsequent analysis. Inappropriate choices of m and Lag can result in the spurious indication of structure in random data (Rapp, et al., 1993). Conversely an inappropriate specification can result in the unnecessary failure to identify structures that are indeed present in the time series. Several candidate criteria for selecting m and Lag have been proposed. An incomplete review of the very large embedding criterion literature is given in Cellucci, et al (2003). Fraser and Swinney (1986) proposed that the best value of Lag to use in an embedding is given by the first minimum of the $I(X_i, X_{i+Lag})$ versus Lag function. This proposal is supported by Abarbanel (1995). To a limited degree the Fraser-Swinney proposal was confirmed in a recent comparative study of embedding criteria (Cellucci, et al., 2003). It should be noted, however, that this comparison was limited to four criteria and is therefore not definitive.

Mutual information calculations are also important in the characterization of causal relationships between two time series. Correlation measures, both linear and nonlinear, quantify the degree of correlation between {X} and {Y} under their respective definitions, but they do not identify causal relationships in the sense of identifying which variable drives the other, if indeed such a relationship exists. The quantification of causal relationships is a problem that is frequently encountered in the investigation of econometric data. Historically the most commonly employed measure of causality in economics research is Granger Causality (Granger, 1969; Kaminski, et al., 2001) which is based on the construction of bivariate autoregressive processes. A complementary procedure for the investigation of causal relationships can be constructed by examining delayed mutual information functions. Stated informally, if a measurement of variable x can predict the future of y more effectively than measurement of y can predict x, then, in that limited sense, in an isolated system variable x can be said to drive variable y. Depending on the complexity of the interacting variables being investigated, causal relationships can be complex functions of time. $I(X_i, Y_{i+\tau})$ is the average number of bits of y at time $t+\tau$ that can be predicted by measuring x at time t. Conversely, $I(Y_i, X_{i+\tau})$ is the average number of bits of x at time $t+\tau$ that can be predicted by measuring y at time t. Xu, et al (1997) describe $I(X_i, Y_{i+\tau})$ as the rate of information transmission from variable x to variable y at a delay of τ. Several investigators have used this technique to assess the time dependence of between channel information transfer in multichannel EEGs (Inouye, et al., 1983, 1993; Lopes da Silva, et al., 1989; Xu, et al., 1997; Chen, et al., 2000).

II. Calculating I(X,Y) with a uniform partition of the XY plane

Let $\{X\} = \{x_1, x_2, x_3 \cdots\cdots x_{N_D}\}$ and $\{Y\} = \{y_1, y_2, y_3 \cdots\cdots y_{N_D}\}$ be time series of equal length. Suppose that the distributions of X and Y, $P_X(i)$ and $P_Y(j)$ are approximated by histograms of N_X and

N_Y elements that uniformly divide the range x_{min} to x_{max} and y_{min} to y_{max}. Though it is commonly the case, it is not necessary for N_X to be equal to N_Y. Indeed, in some instances $N_X = N_Y$ is inappropriate. Suppose that signal X was digitized with 8 bits and that signal Y was digitized with 16 bits. Should this be the case, a Y histogram with a greater resolution is justified. A uniform partition from x_{min} to x_{max} and y_{min} to y_{max} makes the calculation sensitive to outliers. This potentially serious deficiency and the choice of N_X and N_Y will be addressed presently. Let $O_{XY}(i,j)$ denote the occupancy of the (i,j)-th element of the partition of the XY plane that extends from x_{min} to x_{max} on the X axis (N_X equal elements) and from y_{min} to y_{max} on the Y axis (N_Y equal elements). Consider the observed pair (x_k, y_k), $k = 1, \cdots\cdots N_D$. $O_{XY}(i,j)$ is incremented by one if x_k is in the i-th partition element of the X axis and y_k is in the j-th element of the Y axis partition. This process continues for all (x_k, y_k) pairs. $P_{XY}(i,j)$ is determined by normalizing the occupancy against the number of paired observations; $P_{XY}(i,j) = O_{XY}(i,j)/N_D$. The joint probability distribution, $P_{XY}(i,j)$, has $N_X N_Y$ values, many of which may be zero. A discrete approximation of I(X,Y) is computed using the relation derived in Appendix 1.

$$I(X,Y) = \sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} P_{XY}(i,j) \log_2 \left\{ \frac{P_{XY}(i,j)}{P_X(i) P_Y(j)} \right\}$$

where there is no contribution to the sum if $P_{XY}(i,j)$ is equal to zero.

While easy to implement, this procedure for estimating mutual information contains a serious deficiency. The calculation will be sensitive to the choice of N_X and N_Y. An example is shown in the next diagram. $I(X_i, X_{i+Lag})$ is plotted as a function of Lag, for data generated by the Lorenz system.

$$dx/dt = \sigma(x - y)$$

$$dy/dt = -xz + rx - y$$

$$dz/dt = xy - bz$$

where $\sigma = 10$, b=8/3 and r=28. Ten thousand values of the x variable of the Lorenz system were used in calculations where $N_X = N_Y = N_{Elements}$ equally sized elements partition each axis. In these calculations, a well characterized minimum of $I(X_i, X_{i+Lag})$ appears at Lag=18 when $N_{Elements} = 50$. However, as the diagram indicates, this minimum is lost if other values of $N_{Elements}$ are used. Since the location of the first minimum of the $I(X_i, X_{i+Lag})$ versus Lag is frequently the object of a mutual information calculation, this result argues against the common practice of selecting N_X and N_Y

arbitrarily. A rational basis for selecting N_{Elements} can be constructed using a minimum description length argument.

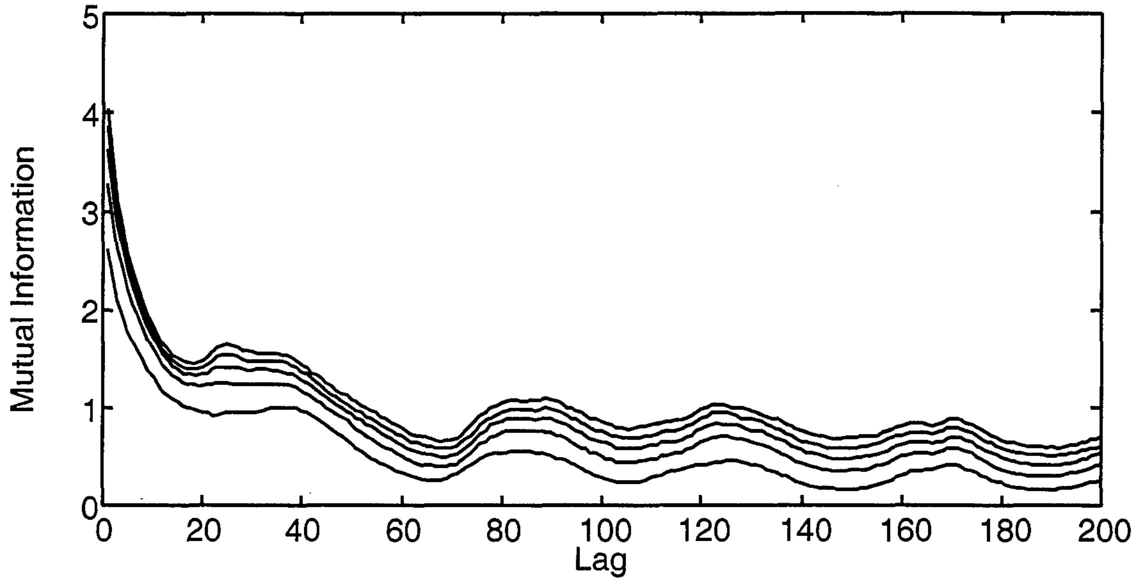

Figure 2. $I(X_i, X_{i+Lag})$ as a function of lag. Ten thousand consecutive values of the Lorenz x variable were used. In the case of the top curve, $N_{\text{Elements}} = 50$. The value of N_{Elements} decrease in steps of ten to the lower curve where $N_{\text{Elements}} = 10$.

The preceding example indicates that the value of mutual information can be sensitive to the number of elements used when a uniform partition of the XY plane is implemented. We must therefore address the question what is the optimal number of elements? This is a restatement of the histogram problem in the specific context of mutual information calculations. The histogram problem is: given a scalar data set $X = \{x_1, x_2, \cdots\cdots x_n\}$, how many elements should be used to construct a histogram of X? If there are too many elements, each element has an occupancy of 0 or 1 and fails to identify the distribution of X in a meaningful way. Similarly, if there are only a small number of elements (consider the limiting case of a single element), the structure of the distribution cannot be discerned. A successful answer therefore lies at an intermediate value. The histogram problem has a long history and has been examined by several investigators (Bendat and Piersol, 1966 page 284; Cocatre-Zilgien and Delcomyn, 1992; Mosteller and Tukey, 1977 page 49).

Tukey suggest that $n^{1/2}$, where n is the number of observations, is the best choice. Bendat and Piersol (1966) recommended $1.87(n-1)^{0.4}$. A systematic theoretical development of the question is given by Rissanen (1989, page 76). Rissanen uses a minimum description length argument to conclude that the optimal value of the number of elements to use in a histogram is the value of m, m_{opt}, that gives a minimum value of the stochastic complexity, F(m).

$$F(m) = n \log_2\left(\frac{R}{m\Delta}\right) + \log_2\binom{n}{n_1, \cdots, n_m} + \log_2\binom{n+m-1}{n}$$

n is the number of data points in set X. R is the range of X, $R = x_{\max} - x_{\min}$. m is the number of elements in a uniform partition, Δ is the resolution of the measurement of x, and $n_1, n_2, \cdots n_m$ are the occupancies of each element in the partition. The multinomial coefficient is

$$\binom{n}{n_1, \cdots, n_m} = \frac{n!}{n_1! n_2! \cdots n_m!}$$

and the binomial coefficient is

$$\binom{n+m-1}{n} = \frac{(n+m-1)!}{n!(m-1)!}$$

The value of Δ only shifts the function by an additive constant. It will not affect the value of m_{opt}. If the only object of the calculation is to determine m_{opt}, Δ can be set equal to 1. Base two logarithms are used throughout the development in Rissanen, but again if the sole object is a determination of m_{opt}, the choice of base is immaterial.

F(M) was calculated using the Lorenz data used to construct Figure 2. A minimum was obtained at $m_{opt} = 32$. Using this value for the number of elements in the uniform partition of the X and Y axes in a calculation of $I(X_i, X_{i+Lag})$ gives a mutual information versus lag function with a well characterized first minimum at Lag=21. This analysis therefore would seem to provide a rational procedure for calculating I(X,Y). Application to the Rössler equations, however, shows the limitations of this approach. The Rössler equations used in the next calculations were

dx/dt=-y-z

dy/dt=x+.2y

dz/dt=.4+xz-5.7z

Using x-axis data generated by this system, a calculation of the Rissanen F(M) gives a minimum at M=40. A forty-element partition of each axis was used in the subsequent calculations of mutual information as a function of Lag for x, y and z-variable data. The resulting mutual information versus Lag functions are shown in Figure 3. It is seen that while x-axis and y-axis data give functions with first minima that are roughly coincident, the function obtained with z-axis data is very different.

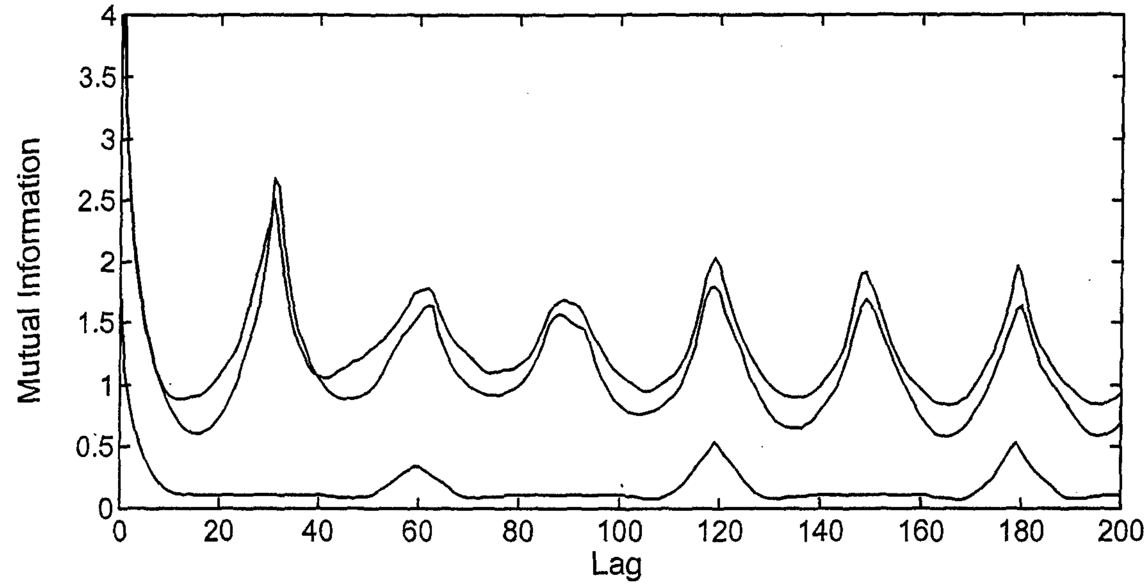

Figure 3. Mutual information $I(X_i, X_{i+Lag})$ as a function of Lag for Rössler data. A uniform partition of the XY plane was constructed using forty elements on each axis. One hundred thousand data points were used. The top curve was obtained with variable x. The curve immediately below it was constructed with variable y data. The lower curve was calculated with variable z data.

The cause of the differences in the z-variable mutual information function in Figure 3 can be identified by examining a three dimensional construction of the trajectory using all three variables (Figure 4). The activity of the Rössler system is confined predominantly to the $z \approx 0$ plane. At irregular, chaotic intervals there is an abrupt excursion into the z>0 domain.

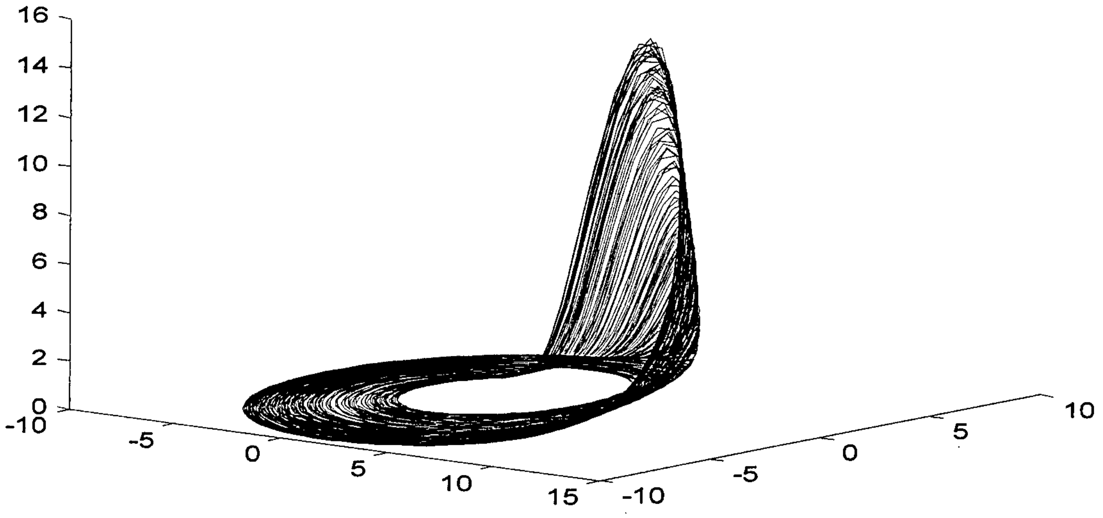

Figure 4. Three-dimensional construction of the Rössler attractor using ten thousand point x, y and z vectors generated using the differential equation and parameter values specified in the text.

An examination of the histograms formed with x, y and z data (Figure 5) shows that while the x and y values are approximately uniformly distributed, most of the activity of the z variable is confined to [0,.375] even though the maximum value of z is approximately 15.

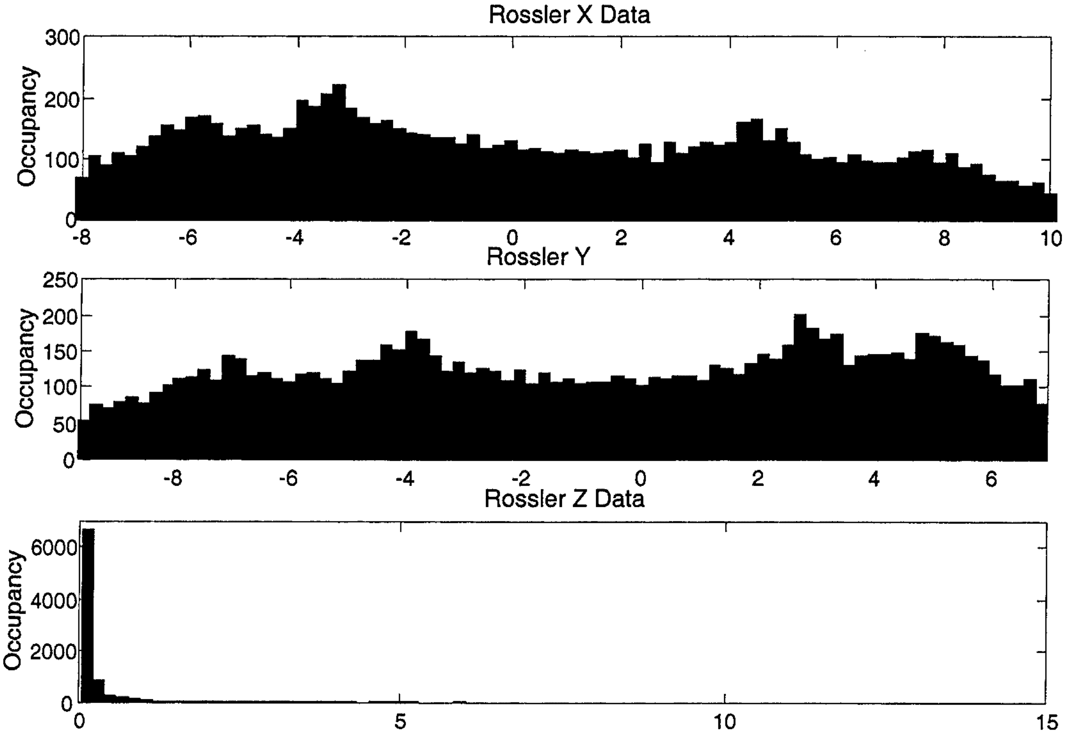

Figure 5. Histograms constructed with Rössler data. The histograms were formed with the ten thousand point vectors used to construct the three dimensional attractor of Figure 4. Note that the ranges of the vertical axes are different.

III. Statistical assessment of I(X,Y) calculations

The results with Rössler data suggest that the calculation of mutual information using a uniform partition can produce misleading conclusions. An alternative to uniform partitioning should, therefore, be sought. An additional and arguably more important issue should also be addressed. The calculations of mutual information should be constructed on a sound statistical foundation. When computing I(X,Y) we should incorporate a statistical test of the confidence of our rejection of the null hypothesis that X and Y are statistically independent. I(X,Y)=0 if X and Y are statistically independent. In practice, we wish to know if a computed nonzero value of I(X,Y) is statistically significant. Therefore, given time series X and Y of length N_D, our object is to assess the null hypothesis that X and Y are statistically independent.

The null hypothesis of statistical independence can be addressed in the following manner. Suppose that the distributions of variables X and Y are approximated by histograms of N_X and N_Y elements. In most applications $N_X = N_Y$, but this is not required. $O_X(i)$ is the observed occupation number of the i-th bin of the variable X histogram. $O_Y(j)$ is assigned analogously. $O_{XY}(i,j)$ is the observed occupation number (a positive integer, not a fractional probability) of element i,j of the XY partition. There are $N_X N_Y$ values of $O_{XY}(i,j)$. Many of them may be zero. $E_{XY}(i,j)$ is the expected occupancy of element i,j of the XY partition given the assumption that X and Y are statistically independent.

$$E_{XY}(i,j) = N_D P_X(i) P_Y(j) = N_D \left\{ \frac{O_X(i)}{N_D} \right\} \left\{ \frac{O_Y(j)}{N_D} \right\} = \frac{O_X(i) O_Y(j)}{N_D}$$

where it should be noted that $E_{XY}(i,j)$ is not necessarily an integer and N_D is the number of x, y pairs.

Following conventional statistical practice (Cochran, 1954; Ott, Longnecker and Ott, 1998), we require $E_{XY}(i,j) \geq 1$ for all elements of the partition and $E_{XY}(i,j) \geq 5$ for at least 80% of these elements. This requirement provides a rational basis for selecting the number of histogram bins N_X and N_Y. If the condition is not met, N_X and N_Y should be decreased. If the condition is satisfied for the initially selected values of N_X and N_Y, then they can be increased and the calculation of $E_{XY}(i,j)$ is repeated. This process is repeated until the highest values of N_X and N_Y consistent with the constraints on $E_{XY}(i,j)$ values are determined. In most applications $N_X = N_Y$, and this calculation is straightforward.

Once the final values of N_X and N_Y are determined and the corresponding values of $O_{XY}(i,j)$ and $E_{XY}(i,j)$ are calculated, the value of χ^2 is calculated.

$$\chi^2 = \sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} \frac{\{O_{XY}(i,j) - E_{XY}(i,j)\}^2}{E_{XY}(i,j)}$$

The condition $E_{XY}(i,j) \geq 1$ for all values of i,j ensures that χ^2 is well behaved. In addition to χ^2, ν, the number of degrees of freedom, is also computed.

$$\nu = (N_X - 1)(N_Y - 1)$$

Using χ^2 and ν, the probability of the statistical independence null hypothesis is computed.

$$P_{\text{Null}} = \text{probability of the null hypothesis} = Q\left(\frac{\nu}{2}, \frac{\chi^2}{2} \right)$$

Q is the incomplete gamma function.

$$Q(x,y)=1-\frac{1}{\Gamma(x)}\int_0^y e^{-t}t^{x-1}dt=\frac{1}{\Gamma(x)}\int_y^\infty e^{-t}t^{x-1}dt \qquad \Gamma(x)=\int_0^\infty e^{-t}t^{x-1}dt$$

IV. Calculation of I(X,Y) using an adaptive XY partition

As previously outlined, we propose that calculation of mutual information should be statistically validated by application of a χ^2 test of the null hypothesis of statistical independence. Additionally, the partition of the XY plane, which is used to calculate the joint probability distribution P_{XY}, should satisfy the Cochran (1954) criterion on the expectancies E_{XY}. Specifically, we require $E_{XY}(i,j) \geq 1$ for all elements of the partition and $E_{XY}(i,j) \geq 5$ for at least 80% of the elements of the partition. In the following algorithm, we use the expectation criterion to construct a nonuniform XY partition. This procedure has two advantages over the use of a naïve uniform partition. First, it reduces sensitivity to outlying values of X and Y. Second, it provides an approximation of the highest partition resolution consistent with the expectation criterion.

Let N_D denote the number of X, Y pairs. N_X is the number of elements used in the partition of the x axis. N_Y is the number of elements used to partition the y axis. For this implementation of the algorithm, N_X and N_Y are equal and denoted by the number of elements N_E. We stress that N_E is the number of elements in the partition of an axis. It is not the number of elements in the XY plane, which is N_E^2. The specification $N_E = N_X = N_Y$ is particularly appropriate when data set Y is a lagged version of data set X. N_E is determined by the following procedure: after determining x_{min} and x_{max}, the x axis is partitioned into N_E elements so that there is an equal occupancy in each element. This partition is nonuniform in the sense that the widths of each element are adjusted individually in order to meet the requirement of uniform occupancy. Let $P_X(i)$ denote the probability of X's membership in the i-th element of the x axis partition. We have

$$P_X(i) = 1/N_E$$

Similarly, after determining y_{min} and y_{max}, the y-axis is partitioned into N_E elements so that there is an equal number of occupants in each y axis element.

$$P_Y(j) = 1/N_E$$

Under the null hypothesis of statistical independence, the expected occupancy of the (i,j)-th element of the partition of the XY plane is

$$E_{XY}(i,j) = N_D P_X(i) P_Y(j) = \frac{N_D}{N_E^2}$$

N_E is determined by finding the largest possible value that gives $E_{XY}(i,j) \geq 5$ for all elements of the XY partition. This criterion is therefore more conservative than the Cochran (1954) criterion that requires E_{XY} to be greater than five in at least 80% of the elements. N_E is the greatest integer such that

$$N_E \leq \left(\frac{N_D}{5} \right)^{1/2}$$

$P_{XY}(i,j)$ is calculated using this partition. Mutual information is calculated with the previously derived formula.

$$I(X,Y) = \sum_{i=1}^{N_E} \sum_{j=1}^{N_E} P_{XY}(i,j) \log \left\{ \frac{P_{XY}(i,j)}{P_X(i) P_Y(j)} \right\}$$

χ^2 and P_{Null} are calculated as previously described. If N_D is exactly divisible by N_E, then the formula for mutual information simplifies and becomes

$$I(X,Y) = \sum_{i=1}^{N_E} \sum_{j=1}^{N_E} P_{XY}(i,j) \log \{ N_E^2 P_{XY}(i,j) \}$$

However, when N_D is not a multiple of N_E, elements of the x axis and y axis partitions do not have exactly identical probabilities equal to $1/N_E$, and the preceding formula should be used. If the Cochran expectation criterion is satisfied (and by construction it will be) and the null hypothesis is not rejected, then, to the extent that can be determined by calculations with this algorithm, the two data sets are statistically independent. Under these conditions, reporting a nonzero value of mutual information cannot be justified. Therefore, in cases where the null hypothesis is not rejected, the algorithm returns I(X,Y)=0 rather than the numerical value produced by the preceding formula.

The application of this procedure to the Rössler data is shown in Figure 6. In contrast with the results of Figure 3, which were obtained with a uniform partition, it is seen that the first minimum of the mutual information versus lag functions obtained with x, y and z-variable data approximately coincide when the adaptive partition is used. The probability of the null hypothesis was calculated for each value of Lag. With these data, P_{Null} was found to be numerically indistinguishable from zero for each value of Lag. Since the data set Y used in these calculations of I(X,Y) is a lagged version of data set X, this rejection of the null hypothesis is anticipated.

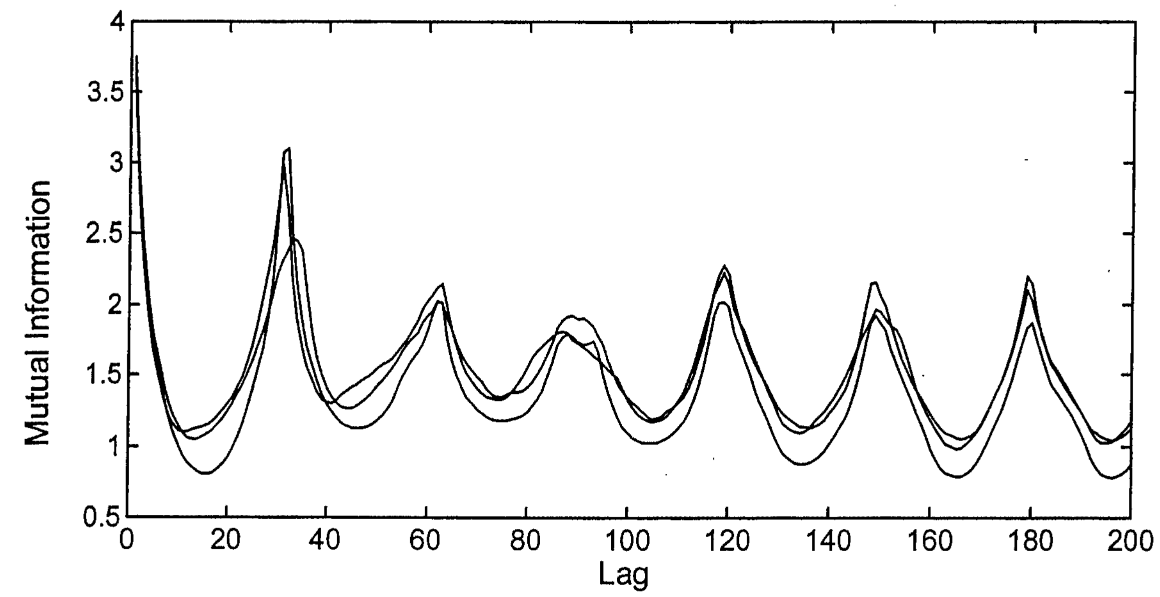

Figure 6. Mutual information as a function of lag using Rössler data. The data used in Figure 3 were used in these calculations. $N_{Data} = 100000$. Viewed at Lag=18, the curves from the x, y and z variables have the top-down order of x to z to y.

In the algorithm constructed here, the number of elements in the X axis partition is equal to the number of elements in the Y-axis partition. This number is denoted by N_E. In this algorithm N_E is determined by the Cochran criterion. Once N_E is specified, the boundaries of the partition's elements are adjusted so that each X-axis element and each y axis element have the same occupancy. $P_{XY}(i, j)$ and I(X,Y) are calculated using this partition. Suppose that time series X is transformed by a monotone increasing function h_X where h_X may be nonlinear. Similarly suppose that time series Y is transformed by a monotone increasing function h_Y. The adaptive partition algorithm for calculating mutual information is then applied to calculate $I(h_X(X), h_Y(Y))$. These transforms are monotonic. Therefore while the values are changed, the relative ordering of elements in the time series are invariant. When the algorithm is applied, the location of the boundaries of axis partitions will be shifted by the occupancies of each element will be unchanged, that is, $P_X(i)$, $P_Y(j)$, and $P_{XY}(i, j)$ are unchanged. Therefore the value of mutual information is unchanged. This is summarized in the following result.

Theorem

Let X and Y be time series of equal length. Let h_X and h_Y be monotone increasing functions. If mutual information is calculated using the adaptive partition algorithm, then

$$I(X,Y) = I(h_X(X), h_Y(Y))$$

Fraser and Swinney (Fraser and Swinney, 1986; Fraser, 1989) have constructed an alternative adaptive partition algorithm. It is described in the next section.

V. The Fraser-Swinney Algorithm

As in the case of the previous algorithm, the calculation is directed to an estimate of the discrete form of the mutual information integral.

$$I(X,Y) = \sum_{i=1}^{N_X}\sum_{j=1}^{N_Y} P_{XY}(i,j)\log\left\{\frac{P_{XY}(i,j)}{P_X(i)P_Y(j)}\right\}$$

Numerical approximation of the joint probability distribution P_{XY} constitutes the most demanding element of the computation. The Fraser-Swinney algorithm (Fraser and Swinney, 1986) does this by constructing a locally adaptive partition of the XY plane.

As a preliminary exercise leading to the construction of the algorithm, consider a sequence of partitions G_0, G_1, G_2, ..., G_m. Each partition is a grid of 4^m elements generated by dividing the X and Y axis into 2^m equiprobable elements, that is the boundaries on the X and Y axis are positioned so that $P_X = P_Y = 1/2^m$ for each element of the partition. G_0 is the entire XY plane. $R_m(K_m)$ denotes an element of the partition G_m. In this notation index K_m runs from 1 to 4^m. $P_{XY}(R_m(K_m))$ is the value of the joint probability distribution on element $R_m(K_m)$.

$$P_{XY}(R_m(K_m)) = N(R_m(K_m))/N_0$$

where $N(R_m(K_m))$ is the occupancy of element $R_m(K_m)$ and N_0 is the total number of XY pairs. Using this notation for the partition and the equiprobability of the X and Y axis partitions gives the following expression for $I_m(X,Y)$, the estimate of mutual information corresponding to partition G_m.

$$I_m(X,Y) = \sum_{K_m=1}^{4^m} P_{XY}(R_m(K_m)) \log\left\{\frac{P_{XY}(R_m(K_m))}{P_X(R_m(K_m))P_Y(R_m(K_m))}\right\}$$

$$I_m(X,Y) = m\log 4 + \sum_{K_m=1}^{4^m} P_{XY}(R_m(K_m)) \log P_{XY}(R_m(K_m))$$

where the first expression for $I_m(X,Y)$ was simplified using the relationships

$$\sum_{K_m=1}^{4^m} P_{XY}(R_m(K_m)) = 1$$

$$P_X(R_m(K_m)) = P_Y(R_m(K_m)) = 1/2^m$$

The essential feature of the Fraser-Swinney algorithm is to take this sequential partitioning procedure and modify it to produce an adaptive partition where the subdivision of any given element is locally determined by the structure of the joint probability distribution P_{XY} on that element. This process can be depicted by the tree structure shown in Figure 7. In this notation, R_0 is the XY plane.

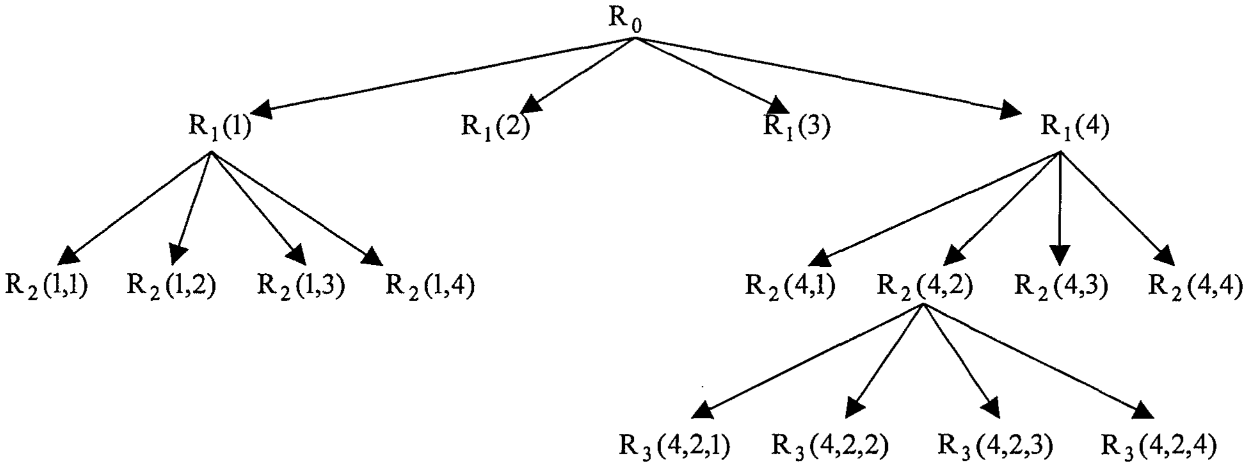

Figure 7. Illustrative example of the adaptive partition employed by the Fraser-Swinney algorithm. In this hypothetical example, the substructure of elements $R_1(2)$ and $R_1(3)$ is approximately uniform and these elements are, therefore, not partitioned. Elements $R_1(1)$, $R_1(4)$ and $R_2(4,2)$ are partitioned into sub-elements because they meet the criterion for the presence of smaller scale structure.

The notation for individual elements of the partition is revised to reflect this structure. As before, each iteration of the partition is effected by a binary equiprobable division of the X and Y axes of an element. In tree notation, an individual element R_m in the partition G_m is identified by an m-tuple $\underline{K}_m = (k_1, k_2, \cdots\cdots k_m)$, $R_m(\underline{K}_m)$.

A finer partition is used in areas of the XY plane where P_{XY} has nonuniform structure. For the hypothetical example in the diagram, P_{XY} is deemed to be approximately uniform on $R_1(2)$ and $R_1(3)$. The partitioning terminates with these elements. In contrast, $R_1(1)$ and $R_1(4)$ have locally nonuniform joint distributions and are partitioned. In this example, partitioning terminates at the G_2 level with the exception of element $R_2(4,2)$, which has a nonuniform joint distribution and is partitioned into four G_3 elements, $R_3(4,2,1)$ through $R_3(4,2,4)$. As previously stated, the partitioning continues until the joint distribution P_{XY} is approximately uniform. A justification for using the uniformity of P_{XY} as the criterion for terminating the partitioning process can be established by examining the special case where P_{XY} is exactly uniform on $R_m(\underline{K}_m)$, where $\underline{K}_m = (k_1, k_2, \cdots k_m)$ is the vector that identifies an element of the order-m partition. P_{XY} is said to be exactly uniform on $R_m(\underline{K}_m)$ if P_{XY} values on the subdivision $R_{m+1}(\underline{K}_m,1)$ to $R_{m+1}(\underline{K}_m,4)$ are equal. For the case of an exactly uniform partition on $R_m(\underline{K}_m)$ we have the following:

$$P_{XY}(R_{m+1}(\underline{K}_m,1)) = P_{XY}(R_{m+1}(\underline{K}_m,2)) = P_{XY}(R_{m+1}(\underline{K}_m,3)) = P_{XY}(R_{m+1}(\underline{K}_m,4))$$

Let $I_m(R_m(\underline{K}_m))$ denote the contribution of element $R_m(\underline{K}_m)$ to mutual information. For the general case, we have

$$I_m(R_m(\underline{K}_m)) = P_{XY}(R_m(\underline{K}_m))\log\left\{\frac{P_{XY}(R_m(\underline{K}_m))}{P_X(R_m(\underline{K}_m))P_Y(R_m(\underline{K}_m))}\right\}$$

where by construction $P_X(R_m(\underline{K}_m)) = P_Y(R_m(\underline{K}_m)) = 1/2^m$. $I_{m+1}(R_m(\underline{K}_m, j))$ is defined analogously on each of the four subdivisions of $R_m(\underline{K}_m)$. Using the equiprobability of the partition gives $P_X(R_{m+1}(\underline{K}_m, j)) = P_Y(R_{m+1}(\underline{K}_m, j)) = 1/2^{m+1}$. The definition of the joint probability distribution gives

$$P_{XY}(R_{m+1}(K_m, j)) = N(R_{m+1}(K_m, j)/N_0$$

where by construction of the partitioning process

$$N(R_m(K_m)) = \sum_{j=1}^{4} N(R_{m+1}(K_m, j))$$

These relationships are generically valid. However, for the special case where P_{XY} is exactly uniform on $R_m(\underline{K}_m)$ it can additionally be shown that

$$I_m(R_m(\underline{K}_m)) = \sum_{j=1}^{4} I_{m+1}(R_{m+1}(\underline{K}_m, j))$$

Thus, in the case where P_{XY} is exactly uniform on $R_m(\underline{K}_m)$, dividing the partition element into four subdivisions will have no effect on the contribution to mutual information obtained from that element. Terminating the partitioning process at level G_m is therefore justified in this case. The qualifying phrase "from that element" is crucial to our understanding of the algorithm. If the uniformity condition were not met, the equality expressed in the immediately preceding equation would not be obtained. As a practical matter, however, it is necessary to establish a criterion that can be used to terminate the partitioning process for some specific element $R_m(\underline{K}_m)$ when P_{XY} is nearly, but not exactly, uniform on that element. In their paper, Fraser and Swinney (1986) construct a test for uniformity that uses a χ^2 test to examine structure on both the m+1 and m+2 generation partition of $R_m(\underline{K}_m)$. Let $N = N(R_m(\underline{K}_m))$ denote the number of XY pairs in element $R_m(\underline{K}_m)$. Using analogous notation for the subdivisions, let $a_i = N(R_{m+1}(\underline{K}_m, i))$ and let $b_{i,j} = N(R_{m+2}(\underline{K}_m, i, j))$. By the Fraser and Swinney criterion, P_{XY} will be deemed to be effectively uniform on $R_m(\underline{K}_m)$ and the partitioning process will be terminated on that element if both $\chi_3^2 < 1.547$ and $\chi_{15}^2 < 1.287$, where

$$\chi_3^2 = \left\{ \frac{16}{9}\left(\frac{1}{N}\right) \sum_{i=1}^{4} (a_i - N/4)^2 \right\}$$

$$\chi_{15}^2 = \left\{ \frac{256}{225}\left(\frac{1}{N}\right) \sum_{i=1}^{4} \sum_{j=1}^{4} (b_{i,j} - N/4)^2 \right\}$$

It should be noted that while the Fraser-Swinney algorithm uses a χ^2 criterion to control subdivisions of the XY plane locally, it does not, in contrast that the algorithm of the previous section, provide a global statistical assessment of an I(X,Y) calculation which includes the probability of the null hypothesis of statistical independence. The code implementing their algorithm distributed by Fraser and Swinney departs from the partition termination criterion outlined in the text of their paper. In their code, the probe for structure is conducted at only one sublevel and the partitioning process is terminated if $\chi_3^2 < 1.547$.

Results obtained when our implementation of the Fraser-Swinney algorithm with a single-level partition termination criterion of $\chi_3^2 < 1.547$ was applied to the Rössler data of Figure 3 are shown in

Figure 8. In our implementation, as in the case of the Fraser-Swinney code, the length of data sets X and Y must be a power of two. Visual comparison of the results obtained with the Fraser-Swinney algorithm and $N_{Data} = 65{,}536$ (Figure 8) with the results obtained with the algorithm of Section IV. and $N_{Data} = 100{,}000$ suggest that similar results were obtained. This point is emphasized in Figure 9 which shows that superposition of the results obtained when $N_{Data} = 65{,}536$ for both algorithms. The values of lag corresponding to the first minimum of the mutual information versus lag function obtained with the two algorithms are either equal or differ by one.

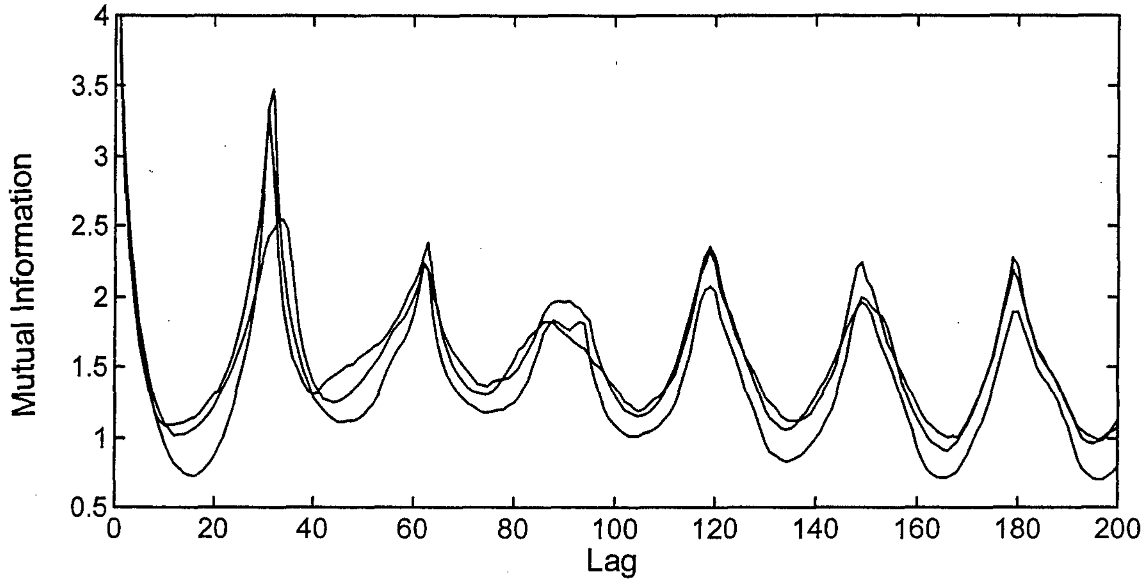

Figure 8. Mutual information as a function of lag using the Rössler data of Figure 3 Mutual information was calculated using the Fraser-Swinney algorithm when $N_D = 65{,}536$. Viewed at Lag=18, the curves from the x, y and z variables have the top-down order of x to z to y.

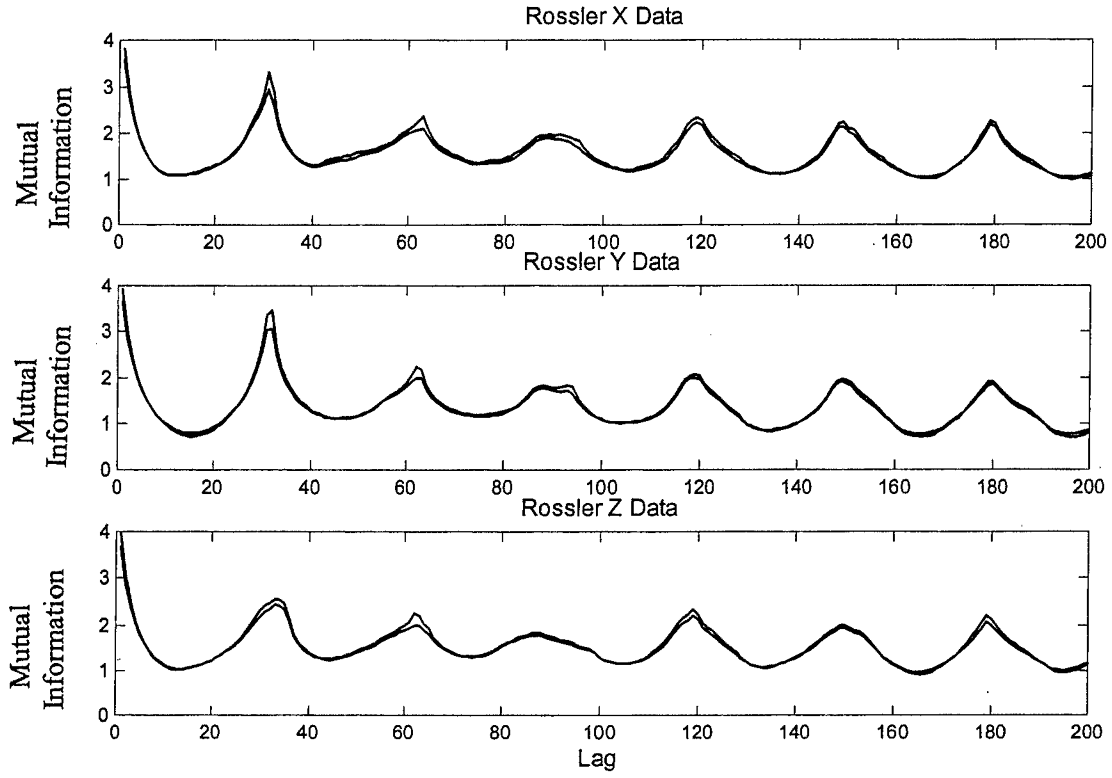

Figure 9. Direct comparison of results obtained with the algorithm of Section IV. and the Fraser-Swinney algorithm using Rössler data of Figure 3. $N_D = 65{,}536$. For those values of lag where the results of the two algorithms differ, the results of the algorithm of Section IV. are below the results obtained with the Fraser-Swinney algorithm

We now have two candidate procedures for calculating I(X,Y), the Fraser-Swinney algorithm and the adaptive partition algorithm presented in Section IV. A procedure for comparing the two methods is constructed in the next section.

VI. Comparing algorithms

In the previous sections, two procedures for computing mutual information were presented. They are compared in this section. Two properties, accuracy and speed, are examined. A comparison of accuracy requires example cases where the true value of mutual information is known to a high accuracy. This can be provided by jointly Gaussian data sets. Two data sets are said to be jointly Gaussian if their joint probability density function centered at (m_x, m_y) has the form

$$P_{XY}(x,y)=\frac{1}{2\pi\sigma_x\sigma_y(1-r^2)^{1/2}}\exp\left\{\frac{-1}{2(1-r^2)}\left[\left(\frac{x-m_x}{\sigma_x}\right)^2-2r\left(\frac{x-m_x}{\sigma_x}\right)\left(\frac{y-m_y}{\sigma_y}\right)+\left(\frac{y-m_y}{\sigma_y}\right)^2\right]\right\}$$

m_x and σ_x are the mean and standard deviation of time series $\{X\}$. m_y and σ_y are defined analogously for $\{Y\}$, and r is the cross-correlation coefficient between $\{X\}$ and $\{Y\}$. For the case of jointly Gaussian data sets, the mutual information is analytically related to the correlation coefficient by

$$I(X,Y)=-0.5\log(1-r^2)$$

(Mars and Lopes da Silva, 1987). A derivation of the relationship is given in Appendix 2. The construction of a procedure for generating jointly Gaussian data sets with a specified correlation coefficient is also presented in that appendix.

Mutual information estimates obtained with the algorithm of Section IV and with the Fraser-Swinney algorithm are compared against $-.5\log(1-r^2)$ for the case of jointly distributed Gaussian data in Figure 9. Ninety nine values of r, uniformly distributed on (-1,1) were used in these calculations. For each value of r, one hundred jointly distributed $\{X\}$, $\{Y\}$ data set pairs of length 8,192 were generated. The average value of mutual information for these pairs was determined using both algorithms. Multiple variants of each algorithm were used. The irregular I(X,Y) versus r function seen in Figure 9 was produced using the Fraser-Swinney algorithm when the sub-partitioning process was terminated with the criterion $\chi_3^2 < 1.547$. With this criterion, an element of the partition is subdivided if the probability of nonuniform substructure is greater than 27%. This is the criterion implemented in their code. Calculations were also performed using $\chi_3^2 < 5.000$. This criterion results in the subdivision of an element of the partition only if the probability of nonuniform substructure is at least 80%. In this case, the results were much closer to $-0.5\log(1-r^2)$. Three variants of the algorithm constructed in Section IV were used. In the first instance, the number of elements in the partition were chosen so that $E_{XY}(i,j)\geq 5$ for all elements. Recall that $E_{XY}(i,j)$ is the expected occupancy in partition element (i,j) given the null hypothesis of statistical independence; $E_{XY}(i,j)=N_D P_X(i) P_Y(j)$ where N_D is the number of elements in the time series $\{X\}$ and $\{Y\}$. Calculations also were performed with the Section IV algorithm with $E_{XY}(i,j)\geq 10$ and with $E_{XY}(i,j)\geq 15$. In the case of the Section IV algorithm, the value I(X,Y)=0 is returned whenever the null hypothesis of statistical independence is not rejected with a confidence level of at least 95%. This convention accounts for the transition to I(X,Y)=0 in the vicinity of r=0 for I(X,Y) functions obtained with this algorithm. Viewed at r=.2 the top-down ordering of the I(X,Y) versus r functions is (i.) Fraser-Swinney algorithm with $\chi_3^2 < 1.547$, (ii.) the algorithm of Section IV with $E_{XY}(i, j) \geq 5$, (iii.) the algorithm of Section IV with $E_{XY}(i, j) \geq 10$, (iv.) the algorithm of Section IV with $E_{XY}(i, j) \geq 15$, (v.) the Fraser-Swinney algorithm with $\chi_3^2 < 5.000$, (vi.) the analytical solution $-0.5\log(1-r^2)$. The greatest numerical value of I(X,Y) is obtained with the Fraser-Swinney algorithm with a subdivision criterion of $\chi^2 < 1.547$ which results in subdivisions whenever the probability of nonuniform substructure exceeds 27%. This produces the greatest value of I(X,Y) because the comparatively tolerant criterion of 27% introduces a numerical indication of small scale structure in the data (and hence a greater value of mutual information) that may not be present. With the more demanding criterion of $\chi^2 < 5.000$, a subdivision is introduced only if there is at least an 80% probability of nonuniform substructure. With this criterion there is less divergence between the algorithm-estimated value of mutual information and the analytically computed value of $-0.5\log(1-r^2)$.

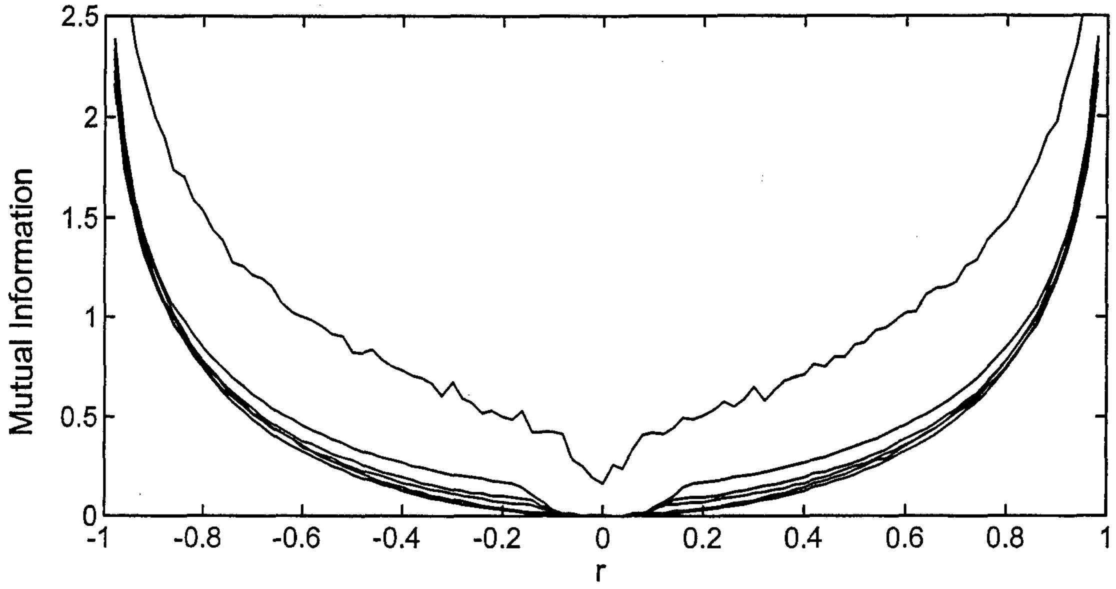

Figure 10. Comparing the Fraser-Swinney algorithm, the algorithm of Section IV. and $-.5\log(1-r^2)$ for jointly distributed Gaussian data. Ninety-nine values of correlation r uniformly distributed on (-1,1) were used. $N_D = 8{,}192$. For each value of r, one hundred {X}, {Y} data set pairs were generated. The algorithm's average value of mutual information is displayed. Viewed at r=.2 the top-down ordering of the I(X,Y) versus r functions is (i.) Fraser-Swinney algorithm with $\chi_3^2 < 1.547$, (ii.) the algorithm of Section IV. with $E_{XY}(i,j) \geq 5$, (iii.) the algorithm of Section IV. with $E_{XY}(i,j) \geq 10$, (iv.) the algorithm of Section IV. with $E_{XY}(i,j) \geq 15$, (v.) the Fraser-Swinney algorithm with $\chi_3^2 < 5.000$, (vi.) the analytical solution $-0.5\log(1-r^2)$

Following Hamilton (1964), the following error measure was calculated.

$$\text{ERROR} = \frac{\sum_{i=1}^{99}\left(I(X,Y)^{\text{Analytical}} - I(X,Y)^{\text{Algorithm}}\right)^2}{\sum_{i=1}^{99}\left(I(X,Y)^{\text{Analytical}}\right)^2}$$

where $I(X,Y)^{\text{Analytical}}$ denotes the value obtained using $-.5\log(1-r^2)$. The results are shown in the next table. It is seen that the magnitude of the error is low with both algorithms.

Table 2. Average Normalized Error in the Estimation of Mutual Information

Algorithm	ERROR
Algorithm of Section IV $E_{XY}(i,j) \geq 5$	1.91×10^{-3}
Algorithm of Section IV $E_{XY}(i,j) \geq 10$	1.55×10^{-3}
Algorithm of Section IV $E_{XY}(i,j) \geq 15$	3.15×10^{-3}
Fraser-Swinney Algorithm $\chi_3^2 < 1.547$	2.48×10^{-1}
Fraser-Swinney Algorithm $\chi_3^2 < 5.000$	0.97×10^{-3}

In addition to providing an explicit assessment of the probability of the null hypothesis of statistical independence, the algorithm of Section IV offers an additional advantage over the Fraser-Swinney algorithm; it is much faster. Comparison of computation times with data sets of different lengths is given in Table 3. Both programs were run in Matlab 6.5.0 (R13) on a Pentium 4 processor running at 2.53 GHz. The computation times of the algorithm of Section IV are typically on the order of .5% of the times required by the Fraser-Swinney Algorithm. In addition to being more accurate than the $\chi_3^2 < 1.547$ criterion, the $\chi_3^2 < 5.000$ algorithm is faster because it introduces fewer subdivisions.

Table 3. Comparative Computation Times for Different Algorithms

N_{Data}	Time Algorithm of Section IV (seconds)	Time Fraser-Swinney Algorithm $\chi_3^2 = 1.547$ (seconds)	Time Fraser-Swinney Algorithm $\chi_3^2 = 5.00$ (seconds)
4096	1.3	266.2	185.2
8192	2.7	544.0	392.4
16384	5.0	1169.5	851.0
32768	9.3	2549.5	1898.5
65536	24.1	5940.5	4533.5

An approximate understanding of the sensitivity of the two algorithms to data set size can be obtained by examining the results presented in Figure 11. That diagram shows the mutual information versus lag functions obtained from a single data set generated by the Rössler equations (x variable data).

As already seen in Figure 8, the results obtained when $N_D = 65536$ are almost identical. More substantive differences are observed, however, when smaller data sets are used. When N_D is 4096 and 8192, the algorithm of Section IV produces output that is slightly less than, but largely parallel to, the results obtained when $N_D = 65536$. For this algorithm, the value of Lag giving the first minimum of mutual information was the same for all values of N_D tested. In contrast, when $N_D = 4096$ and 8192, the Fraser-Swinney algorithm produces mutual information versus lag functions that present structures that are lost when more data are incorporated into the computations. In some instances, these structures can alter the identification of the lag giving the minimum value of mutual information.

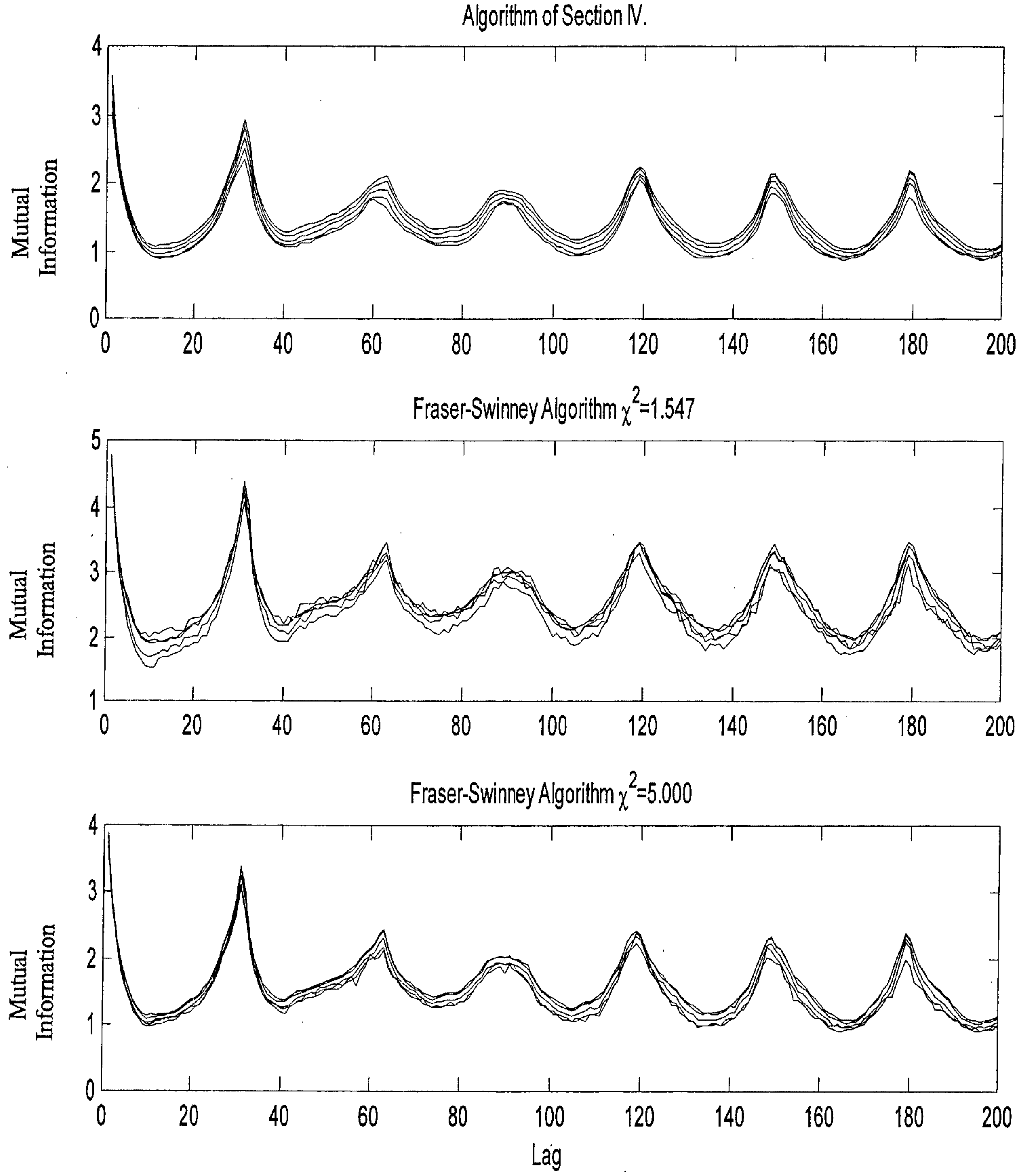

Figure 11. Mutual Information versus Lag for Data Sets of Different Sizes. Mutual information versus lag was computed using both algorithms for $N_D = 4096$, 8192, 16384, 32768 and 65536. The data were generated by the Rössler equations, and x-variable output was used in the calculations. Functions calculated with $N_D = 65536$ are at the top of each set of curves. Functions calculated with $N_D = 4096$ are at the bottom of each set of curves.

VII. Discussion

The Fraser-Swinney algorithm with the $\chi_3^2 < 5.000$ criterion out-performs that algorithm when $\chi_3^2 < 1.547$ is used both in terms of accuracy (Table 2) and speed (Table 3). Given a dichotomous choice between these two options, the $\chi_3^2 < 5.000$ variant would be preferable. A comparison of the Fraser-Swinney algorithm with the $\chi_3^2 < 5.000$ criterion against the algorithm of Section IV leads to the following conclusions. First, the algorithm of Section IV has a significant advantage over the Fraser-Swinney algorithm in providing a global test of the statistical independence null hypothesis. The Fraser-Swinney algorithm uses a χ^2 test locally to implement the partitioning protocol. It does not, however, return an assessment of the statistical independence of X and Y. Second, while the Fraser-Swinney algorithm is more accurate with data sets where $N_D = 8192$ (Table 2), the results of Figure 11 suggest that the Fraser-Swinney algorithm requires large data sets even when the $\chi_3^2 < 5.000$ criterion is used. When smaller data sets are used the Fraser-Swinney algorithm presents structures that disappear when more data becomes available. If the object of the calculation is to use $I(x_i, x_{i+Lag})$ functions to find the appropriate Lag for embedding, then these local minima could give misleading results. Third, the algorithm of Section IV requires about .5% of the calculation time required by the Fraser-Swinney algorithm.

Limitations of this study should be noted. Additional algorithms could be considered. Following Silverman (1986), Moon, et al. (1995) have used kernel density estimators to calculate probability densities. They argue that the resulting algorithm outperforms the Fraser-Swinney algorithm. Moon, et al. also suggest that their algorithm can be improved by using K-d trees to partition the data. Caution must be exercised when evaluating this suggestion. Our exploratory calculations have shown that K-d tree partitions can be very sensitive to initial conditions. This sensitivity is addressed by Bradley and Fayyad (1998) who published a procedure for computing initial conditions based on a procedure for estimating the modes of a distribution.

Instead of partitioning phase space as is done in the algorithms discussed above, Pawelzik and Schuster (1987) used the first order correlation integral to calculate probability densities and entropies. These entropies are then used to calculate mutual information. We consider here application of the technique to embedded time series data, $X_k = (x_k, x_{k+Lag}, x_{k+2Lag}, \cdots\cdots x_{k+(m-1)Lag})$ and

$Y_k = (y_k, y_{k+Lag}, y_{k+2Lag}, \cdots\cdots y_{k+(m-1)Lag})$ k = 1, ..., N-m+1. Application to scalar data is trivially obtained by taking the embedding dimension, m, to be one for X and Y, and thus dimension two for the joint space. The density of X in the neighborhood of X_k is approximated by the first order correlation integral,

$$p_{X_k}(r) = \frac{1}{N_V - 1}\sum_{j \neq k} \Theta\left(r - |X_j - X_k| \right),$$

where Θ is the Heaviside function, N_V is the number of embedding vectors, and r is the neighborhood size being considered. This density differs from that used earlier because it counts the number of points in possibly overlapping neighborhoods. The densities used in the algorithms discussed earlier involved non-overlapping neighborhoods created by the partitioning process. This leads to a slightly different expression for the entropy which, in this case, is given by

$$H(X,r) = -\frac{1}{N_V}\sum_{k=1}^{N} \ln p_{X_k}(r)$$

In some implementations, finite sample corrections due to Grassberger are included (Grassberger, 1988). The entropies of the Y data as well as the joint entropy are calculated similarly, and these are used to obtain the mutual information from the relation I(X,Y)=H(X)+H(Y)-H(X,Y).

Quian Quiroga et al., (2002) used the Pawelzik-Schuster algorithm with the Grassberger corrections in a study of synchronization of rat electrocorticograms. They studied three multichannel ECoG records in a rat model of genetic absence epilepsy and compared activity between left and right hemispheres. (The data they analyzed are available at www.vis.caltech.edu/~rodri). The first record, their example A, was obtained in an interictal condition and the remaining two, their examples B and C, were recorded during seizures and presented repetitive spike discharges. In addition to mutual information, the synchronization measures examined included nonlinear interdependencies, phase synchronizations, cross correlation, and the coherence function. They concluded that except for mutual information their linear and nonlinear measures provided qualitatively similar results. Namely, interhemispheric synchronization was highest in example B, followed by A and then C. The authors felt that the small number of data points (N=1000) was responsible for the failure of mutual information to provide robust estimates of interhemispheric synchronization.

These data were re-analyzed by Duckrow and Albano (2003) using a modified Fraser-Swinney algorithm. The data were embedded and interleaved as described in Appendix 3 and the resulting binary representations were used as inputs in the Fraser Swinney algorithm. Using embedding dimensions from

1 to 10 and Lags from 1 to 30, the results consistently showed the B > A > C ranking that Quian Quiroga *et al.* found using other measures of synchronization. Results obtained by Duckrow and Albano using these data and a uniform partition algorithm showed a behavior similar to that found by Quian Quiroga when they used the Pawelzik-Schuster algorithm. The expected ranking of B >A > C was seen only for embedding dimensions 1 and 2, and the value of mutual information increased with increasing embedding dimension (Figure 12). This is likely a consequence of small data set size. An examination of the uniform partition algorithm indicates why this should be the case. If one considers X and Y as N uniformly distributed random numbers, the individual and joint probabilities would be $P_X = 1/N$, $P_Y = 1/N$, and $P_{XY} = 1/N^2$. The resulting mutual information would be zero. However, when sparse data are scattered into a large multidimensional embedding space it becomes unlikely that more than one point is present in any one histogram element. In this case the joint probability is closer to 1/N than $1/N^2$. The resulting mutual information would be log(N). When a fixed bin-width histogram method was applied to sets of N=1000 embedded uniformly distributed random numbers, Duckrow and Albano found that the resulting average mutual information value rose quickly with increasing embedding dimension and rapidly approached log(N), as shown by the dash-dot line in Fig. 12.

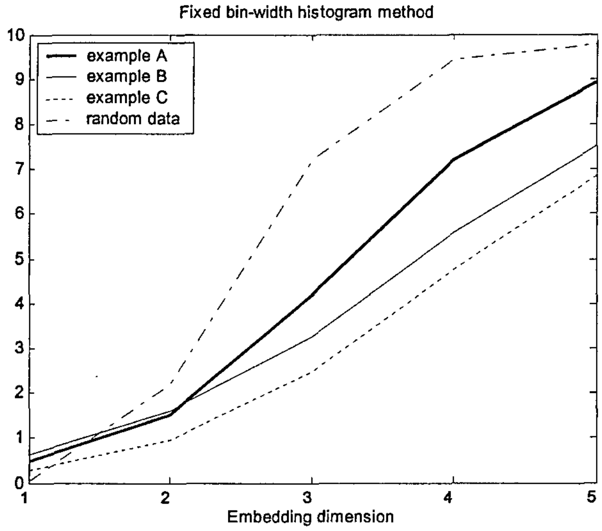

Figure 12. Average mutual information at increasing embedding dimension with fixed Lag = 10 (From Duckrow and Albano, 2003).

In contrast, Albano and Duckrow found that the average mutual information of same the random numbers calculated with the Fraser-Swinney algorithm had a median value of 1.1 × 10^{-5} and did not exceed 0.02. Interleaving data embedded in m dimensions represents m-dimensional vectors as scalars and reduces the calculation of the joint probability in a 2m-dimensional space to a two-dimensional calculation. Increasing the embedding dimension does not make the embedded points any sparser in the reduced space, and, therefore, calculation of the mutual information proceeds in the same manner as for calculation for unembedded data. One does not get the artificial increase in the calculated mutual information that results from scattering a limited number of points in spaces of ever-increasing dimensions observed with uniform partition algorithms. In subsequent work, Quian Quiroga et al. (2003) repeated their calculations using interleaved embedded data as inputs to the Pawelzik-Schuster algorithm and confirmed the B > A > C ranking of mutual information reported by Albano and Duckrow. In this contribution, they suggest that a more precise method for calculating the mutual information may be one that estimates probability densities using k-th nearest neighbor distances rather than the number of neighbors in neighborhoods of fixed size.

Yet another approach to calculating mutual information has been published by Kilminster, et al. (2002) who have shown that the Radon transform can be used to estimate joint probability density functions which can then be used to estimate mutual information. They argue that, in contrast with standard methods, this procedure preserves fractal structure. The Kilminster et al. and the Moon, et al. algorithms could be compared against the Fraser-Swinney algorithm and the algorithm of Section IV in an expanded study.

Acknowledgements

We would like to thank E. Kurali for directing us to information concerning jointly Gaussian data sets. The comments and discussions with A. I. Mees and T. I. Schmah are acknowledged with gratitude. Particular thanks are directed to CDR R. S. Hernandez, USN for encouragement, support and project leadership. This research was supported by Grant Number 601135N.4508.518.A0247 from the Office of Naval Research and the Navy Bureau of Medicine to the Naval Medical Research Center. The opinions expressed here are those of the authors and do not reflect the official policy or position of the Department of the Navy or the U.S. Government.

Bibliography

Bendat,J.B. and Piersol,A.G. (1966). Measurement and Analysis of Random Data. John Wiley, NY.

Bradley,P.S. and Fayyad,U.M. (1998). Refining initial points for k-means clustering. In: Proceedings of the Fifteenth Conference on Machine Learning. I Brasko and S. Dzeroski, eds. Morgan Kaufman, New York.

Callaway,E. and Harris,P.R. (1974). Coupling between cortical potentials from different areas. Science. 183, 873-875.

Cellucci,C.J., Albano,A.M. and Rapp,P.E. (2003). Comparative study of embedding methods. Physical Review E. 67, 066210-1 to 066210-13.

Chen,F., Xu,J., Gu,F., Yu,X., Meng,X. and Qiu,Z. (2000). Dynamic processes of information transmission complexity in human brains. Biological Cybernetics. 83(4), 355-366.

Cocatre-Zilgien,J.H. and Delcomyn,F. (1992). Identification of bursts in spike trains. J. Neurosci. Methods. 41, 19-30.

Cochran,W.G. (1954). Some methods for strengthening the common χ^2 test. Biometrics. 10, 417-451.

Cover,T.M. and Thomas,J.A. (1991). Elements of Information Theory. Wiley, NY.

Duckrow,R.B. and Albano,A.M. (2003). Comment on "Performance of different synchronization measures in real data: A case study on electroencephalographic signals." Physical Review E. 67, 063901-1 to 063901-3

Fraser,A.M. (1989). Information storage and entropy in strange attractors. IEEE Trans. Inform. Theory. 35, 245-262.

Fraser,A.M. and Swinney,H.L. (1986). Independent coordinates for strange attractors from mutual information. Phys. Rev. 33A, 1134-1140.

Granger,C.W.J. (1969). Investigating causal relations by econometric models and cross-spectral methods. Econometrica. 37, 424-438.

Grassberger,P. (1988). Finite sample corrections to entropy and dimension estimates. Physics Letters A. 128, 369.

Hamilton,J.D. (1964). Time Series Analysis. Princeton University Press, Princeton, NJ.

Inouye,T., Shinosaki,K. and Yagasaki,A. (1983). The direction of spread of alpha activity over the scalp. Electroenceph. clin. Neurophysiol. 55, 290-300.

Inouye,T., Shinosaki,K., Iyama,A. and Matsumoto,Y. (1993). Localization of activated areas and directional EEG patterns during mental arithmetic. Electroencephalography and Clinical Neurophysiol. 86, 224-230.

Kaminski,M.J., Ding,M., Truccolo,W.A. and Bressler,S.L. (2001). Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance. Biological Cybernetics. 85, 145-157.

Kilminster,D., Allingham,D. and Mees,A. (2002). Estimating invariant probability densities for dynamical systems. Annals of the Institute of Statistical Mathematics. 54(1), 224-233.

Lopes da Silva, F., Pijn,J.P. and Boeijinga,P. (1989). Interdependence of EEG signals: Linear vs. nonlinear associations and the significance of time delays and phase shifts. Brain Topography. 2, 9-18.

Mañé,R. (1980). On the dimension of the compact invariant sets of certain nonlinear maps. In: Dynamical Systems and Turbulence. Lecture Notes in Mathematics. Volume 898. D.A.Rand and L.S.Young, eds. pp. 230-242. Springer-Verlag, NY.

Mars,N.J.I., Thompson,P.M. and Wilkus,R.J. (1985). Spread of epileptic seizure activity in humans. Epilepsia. 26, 85-94.

Mars,N.J.I. and Lopes da Silva,F.H. (1987). EEG analysis methods based on information theory. In: Methods of Analysis of Brain Electrical and Magnetic Signals. EEG Handbook (Revised Series, Volume 1). A. S. Gevins and A. Rémond, eds. pp. 297-307. Amsterdam: Elsevier Science Publishers.

Moon,Y.-I., Rajagopalan,R. and Lall,U. (1995). Estimation of mutual information using kernel density estimators. Physical Review E., 52, 2318-2321.

Mosteller,F. and Tukey,J.W. (1977). Data Analysis and Regression. Addison-Wesley, Reading, MA.

Ott,L., Longnecker,M.T. and Ott,R.L. (1998). An Introduction to Statistical Methods and Data Analysis. New York: Wadsworth.

Pawelzik,K. and Schuster,H.G. (1987). Generalized dimensions and entropies from a measured time series. Physical Review A. 35, 481-484.

Press,W.H., Flannery,B.P., Teukolsky,S.A. and Vetterling,W.T. (1986). Numerical Recipes. The Art of Scientific Computing. Cambridge University Press, Cambridge.

Quian Quiroga,R., Kraskov,A., Kreuz,T. and Grassberger,P. (2002). Performance of different synchronization measures in real data: A case study on electroencephalographic signals. Physical Review. 65E, 041903-1 to 041903-14.

Quian Quiroga,R., Kraskov,A., Kreuz,T. and Grassberger,P. (2003). Reply to "Comments on 'Performance of different synchronization measures in real data: A case study on electroencephalographic signals'." Physical Review E 67, 063902-1 to 063902-2.

Rapp,P.E., Albano,A.M., Schmah,T.I. and Farwell,L.A. (1993). Filtered noise can mimic low dimensional chaotic attractors. Phys. Rev. 47E, 2289-2297.

Rissanen,Y. (1992). Stochastic Complexity in Statistical Inquiry. World Scientific, Singapore.

Silverman,B.W. (1986). Density Estimation for Statistics and Data Analysis. New York: Chapman and Hall.

Takens,F. (1980). Detecting strange attractors in turbulence. Lecture Notes in Mathematics. Volume 898. D.A.Rand and L.S.Young, eds. pp. 365-381. Springer-Verlag, NY.

Xu,J., Liu,Z.-R., Liu,R. and Yang,Q.-F. (1997). The information transmission of human brain cortex. Physica. 106D, 363-374.

Appendix 1. Mutual Information: Definition and Mathematical Characterization

A. Define a system, $\{x_1, x_2, \cdots\cdots x_{N_X}\}$, $\{P_X(1), P_X(2), \cdots\cdots P_X(N_X)\}$

B. Define the information in the i-th symbol, $I(i)$

C. Define the entropy of a system, H(X)

D. Define the joint probability distribution, $P_{XY}(i, j)$

E. Define the conditional probability distribution, $P_{X|Y}(i, j)$

F. Define the joint entropy, H(X,Y)

G. Define the conditional entropy H(X|Y)

H. Define mutual information I(X,Y)

A. Define a system, $\{x_1, x_2, \cdots\cdots x_{N_X}\}$, $\{P_X(1), P_X(2), \cdots\cdots P_X(N_X)\}$

X is a system consisting of a discrete set of possible symbols, $\{x_1, x_2, \cdots\cdots x_{N_X}\}$. The set of possible symbols, which is often referred to as the alphabet, is to be distinguished from an output sequence, or message, generated by system X. An output sequence is an ordered sequence of symbols drawn from the symbol set. Throughout we use N_X to denote the number of elements in the symbol set and N_D, where 'D' denotes data, to denote the length of a message. The associated probabilities of each element of the symbol set is given by $\{P_X(1), P_X(2), \cdots\cdots P_X(N_X)\}$, where probabilities have the property

$$\sum_{i=1}^{N_X} P_X(i) = 1$$

In many of the applications implemented here, symbol x_i denotes the presence of an event in the i-th element of a histogram that is composed of N_X bins. In this context, $P_X(i)$ is the occupation probability of the i-th bin.

B. Define the information in a symbol, $I(i)$

The information content of the i-th symbol is

$$I(i) = -\log_2 P_X(i)$$

Throughout, all logarithms are computed in base 2, and the resulting values are reported in bits. (In some texts the natural logarithm is used and the reported units are "nats.") Suppose the probability of a symbol is 1. In this case, nothing is learned by observing it. The corresponding information is $-\log_2(1) = 0$. As a symbol become increasingly improbable, its associated information increases.

C. Define the entropy of a system, H(X)

Information is a property of a symbol. In contrast, entropy is a property of a system. Entropy, H(X), is the average amount of information gained from an observation of x. Restated, H(X) is the average uncertainty in x prior to its observation.

$$H(X)=\sum_{i=1}^{N_X} P_X(i)I(i)$$

$$H(X)=-\sum_{i=1}^{N_X} P_X(i)\log_2 P_X(i)$$

This can be generalized to a continuous variable

$$H(X)=-\int P_X(x)\log_2 P_X(x)dx$$

In the discrete case, suppose an observation can be in one of sixty-four bins. Thus there are sixty-four possible symbols, $\{x_1,\cdots\cdots x_{64}\}$. Suppose further that the probability of any given symbol is the same, $P(i)=1/64$ for all i.

$$H(X)=-\sum_{i=1}^{64} P_X(i)\log_2 P_X(i)=-64\left(\frac{1}{64}\right)\log_2\left(\frac{1}{64}\right)=\log_2 2^6=6 \text{ bits}$$

Similarly, if there were 128 equiprobable symbols, H(X)=7 bits.

For the general case, using the convexity of xlogx, it can be shown that the maximum value of entropy is obtained when all symbols have the same probability. Additionally, using a series expansion of logx, it can be shown that

$$\lim_{x\to 0} x\log x=0$$

Therefore, as previously asserted, if $P_X(j)=1$ and $P_X(k)=0$ for all $k\neq j$, then H(X)=0.

D. Define the joint probability distribution, $P_{XY}(i,j)$

Consider two systems, X with the output sequence $\{x_1,x_2,\cdots\cdots x_{N_D}\}$ and system Y presenting the message $\{y_1,y_2,\cdots\cdots y_{N_D}\}$. When defining a joint probability distribution, one must emphasize the previously made distinction between N_X the number of different symbols that can be presented by system X, N_Y the number of distinct symbols that can be presented by system Y, and N_D the number of observed (x,y) pairs. As before, the independent probability of each symbol of system X is denoted by

$\{P_X(1), P_X(2), \cdots\cdots P_X(N_X)\}$. Similarly, the independent probability distribution of system Y is $\{P_Y(1), P_Y(2), \cdots\cdots P_Y(N_Y)\}$. The joint probability distribution $P_{XY}(i,j)$ is the probability that an (x,y) pair consists of the i-th system X symbol and the j-th system Y symbol. It should be noted that $P_{XY}(i,j) = P_X(i)P_Y(j)$ if and only if X and Y are independent.

Lemma: Relationship between joint probability distributions and single variable distributions

$$P_X(i) = \sum_{j=1}^{N_Y} P_{XY}(i,j)$$

Demonstration:

By summing over all possible y values, the probability of y, whatever value it might be, is 1. Therefore the remaining value in the sum is the probability of the system X symbol.

E. Define the conditional probability distribution, $P_{X|Y}(i,j)$

Given systems X and Y defined above, the conditional probability distribution is denoted by $P_{X|Y}$. $P_{X|Y}(i,j)$ is the probability that $X = x_i$ given that it is already known that $Y = y_j$. The relationship between the joint probability distribution and the conditional probability distribution is established by the following lemma.

Lemma: The relationship between the conditional probability distribution and joint probability distribution is given by:

$$P_{XY}(i,j) = P_{X|Y}(i,j)P_Y(j)$$

Demonstration:

$P_Y(j)$ is the probability that $Y = y_j$. $P_{X|Y}(i,j)$ is the probability that $X = x_i$ if it is already known that $Y = y_j$. Therefore the product of these probabilities is the probability that both $X = x_i$ and $Y = y_j$, which by definition is $P_{XY}(i,j)$.

If X and Y are independent, then $P_{XY}(i,j) = P_X(i)P_Y(j)$, and $P_X(i) = P_{X|Y}(i,j)$. That is, if X and Y are independent, then the probability that $X = x_i$ is determined solely by system X.

F. Define the joint entropy, H(X,Y)

Given X and Y systems as previously defined, the joint entropy is defined as:

$$H(X,Y) = -\sum_{i=1}^{N_X}\sum_{j=1}^{N_Y} P_{XY}(i,j)\log_2 P_{XY}(i,j)$$

H(X,Y) is the average amount of information gained by observing an (x,y) pair.

Lemma: The joint entropy of a system with itself is given by

$$H(X,X)=H(X)$$

Demonstration:

By definition

$$H(X,Y) = -\sum_{i=1}^{N_X}\sum_{j=1}^{N_Y} P_{XY}(i,j)\log_2 P_{XY}(i,j)$$

If Y=X, then $P_{XY}(i,j) = \delta_{ij}P_X(i)$ where δ_{ij} is Kronecker's delta

$$H(X,Y) = -\sum_{i=1}^{N_X}\sum_{j=1}^{N_Y} \delta_{ij}P_X(i)\log_2 \delta_{ij}P_X(i)$$

$$\lim_{z\to 0} z\log z = 0$$

Hence

$$H(X,X) = -\sum_{i=1}^{N_X} P_X(i)\log_2 P_X(i) = H(X)$$

Lemma: The joint entropy function is symmetric, that is

$$H(X,Y)=H(Y,X)$$

Demonstration:

By definition

$$H(X,Y) = -\sum_{i=1}^{N_X}\sum_{j=1}^{N_Y} P_{XY}(i,j)\log_2 P_{XY}(i,j)$$

The joint probability distribution is not a conditional probability distribution; $P_{XY}(i,j) = P_{YX}(j,i)$. Therefore:

$$H(X,Y)=H(Y,X)$$

G. Define the conditional entropy H(X|Y)

For a given value of Y, say $Y = y_j$, the conditional entropy is defined as

$$H(X \mid j) = -\sum_{i=1}^{N_X} P_{X|Y}(i,j)\log_2 P_{X|Y}(i,j)$$

$H(X \mid j)$ is the average amount of information obtained by observing X when $Y = y_j$. H(X|Y), the average conditional entropy, is $H(X \mid j)$ averaged over all possible y's.

$$H(X \mid Y) = \sum_{j=1}^{N_Y} P_Y(j) H(X \mid j)$$

H(X|Y) is the average information obtained by observing X after Y is known. Said another way, H(X|Y) is the average number of additional bits required to specify X if Y is known.

Lemma: Relationship between conditional entropy and joint entropy is given by

$$H(X|Y)=H(X,Y)-H(Y)$$

Demonstration:

By definition

$$H(X \mid Y) = \sum_{j=1}^{N_Y} P_Y(j) H(X \mid j)$$

where

$$H(X \mid j) = -\sum_{i=1}^{N_X} P_{X|Y}(i,j) \log_2 P_{X|Y}(i,j)$$

It has been demonstrated previously that

$$P_{X|Y}(i,j) P_Y(j) = P_{XY}(i,j)$$

Therefore

$$H(X \mid Y) = \sum_{j=1}^{N_Y} P_Y(j)(-1) \sum_{i=1}^{N_X} \frac{P_{XY}(i,j)}{P_Y(j)} \log_2 \frac{P_{XY}(i,j)}{P_Y(j)}$$

$$H(X \mid Y) = -\sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} P_{XY}(i,j) \log_2 \frac{P_{XY}(i,j)}{P_Y(j)}$$

$$H(X \mid Y) = -\sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} P_{XY}(i,j) \log_2 P_{XY}(i,j) \quad + \quad \sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} P_{XY}(i,j) \log_2 P_Y(j)$$

The first term on the right hand side is the joint entropy.

$$H(X \mid Y) = H(X,Y) + \sum_{j=1}^{N_Y} \left\{ \sum_{i=1}^{N_X} P_{XY}(i,j) \right\} \log_2 P_Y(j)$$

It was previously shown that

$$\sum_{i=1}^{N_X} P_{XY}(i,j) = P_Y(j)$$

Hence

$$H(X \mid Y) = H(X,Y) + \sum_{j=1}^{N_Y} P_Y(j) \log_2 P(j)$$

$$H(X|Y)=H(X,Y)-H(Y)$$

H. Define mutual information I(X,Y)

The average amount of information obtained by observing X can be conceptualized as consisting of two components.

Average amount of information obtained by an observation of X

= Average amount of information about X obtained by observing Y

+ Average amount of information about X obtained by observing X after Y is known

Two of these elements have already been defined.

H(X) = Average amount of information obtained by an observation of X

H(X|Y) = Average amount of information about X obtained by observing X after Y is known

The third element is defined as the mutual information

I(X,Y) = Average amount of information about X obtained by observing Y

$$H(X) = I(X,Y)+H(X|Y)$$

$$I(X,Y)=H(X)-H(X|Y)$$

Shannon (1948) used the phrase “rate of transmission” for this quantity

Properties of Mutual Information

Lemma 1. H(X)=I(X,X), referred to as the self-information

Lemma 2. I(X,Y)=H(X)+H(Y)-H(X,Y), note that this is the joint entropy, not the conditional entropy

Lemma 3. I(X,Y)=I(Y,X), mutual information is symmetric

Lemma 4. I(X,Y)=H(Y)-H(Y|X)

Lemma 5. $I(X,Y) = \sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} P_{XY}(i,j) \log_2 \left\{ \frac{P_{XY}(i,j)}{P_X(i) P_Y(j)} \right\}$ This is the expression that will be used in most of the computations.

Lemma 6. I(X,Y)=0 if X and Y are independent

Lemma 1. The relationship between entropy H(X) and self-information I(X,X) is given by

$$H(X)=I(X,X)$$

Demonstration:

By definition

$$I(X,Y)=H(X)-H(X|Y)$$

It was previously demonstrated that

$$H(X|Y)=H(X,Y)-H(Y), \text{ and}$$
$$H(X,X)=H(X)$$

Hence

$$\begin{aligned} I(X,X) &= H(X)-H(X|X) \\ &= H(X)-\{H(X,X)-H(X)\} \\ &= H(X)-\{H(X)-H(X)\} \\ &= H(X) \end{aligned}$$

Lemma 2. Mutual information I(X,Y) can be expressed in terms of entropies H(X), H(Y) and joint entropy H(X,Y) by

$$I(X,Y)=H(X)+H(Y)-H(X,Y)$$

Demonstration:

By definition

$$I(X,Y)=H(X)-H(X|Y)$$

From a previous result

$$H(X|Y)=H(X,Y)-H(Y)$$

Therefore:

$$I(X,Y)=H(X)+H(Y)-H(X,Y)$$

Lemma 3. Mutual information is symmetrical, that is

$$I(X,Y)=I(Y,X)$$

Demonstration:

By Lemma 2,

$$I(X,Y)=H(X)+H(Y)-H(X,Y)$$

Therefore:

$$I(Y,X)=H(Y)+H(X)-H(Y,X)$$

It was previously shown that the joint entropy is symmetric. This gives

$$I(Y,X)=H(Y)+H(X)-H(X,Y)=I(X,Y)$$

Lemma 4. The relationship between mutual information and joint entropy is given by

$$I(X,Y)=H(Y)-H(Y|X)$$

By Lemma 2, the left hand side of the equation is:

$$LHS=H(X)+H(Y)-H(X,Y)$$

By a previous lemma,

$$H(Y|X)=H(Y,X)-H(X)$$

Using this and the symmetry of the joint entropy gives the following expression for the right hand side of the equation

$$\begin{aligned}RHS&=H(Y)-H(H|X)\\&=H(Y)-\{H(X,Y)-H(X)\}\\&=H(X)+H(Y)-H(X,Y)=I(X,Y)\end{aligned}$$

Lemma 5. Computational Expression: Mutual information can be expressed in terms of probability and joint probability distributions by

$$I(X,Y)=\sum_{i=1}^{N_X}\sum_{j=1}^{N_Y}P_{XY}(i,j)\log_2\left\{\frac{P_{XY}(i,j)}{P_X(i)P_Y(j)}\right\}$$

Demonstration:

Expanding the right hand side of this expression gives:

$$\begin{aligned}RHS=&\sum_{i=1}^{N_X}\sum_{j=1}^{N_Y}P_{XY}(i,j)\log_2 P_{XY}(i,j)\\&-\sum_{i=1}^{N_X}\sum_{j=1}^{N_Y}P_{XY}(i,j)\log_2 P_X(i)\\&-\sum_{i=1}^{N_X}\sum_{j=1}^{N_Y}P_{XY}(i,j)\log_2 P_Y(j)\end{aligned}$$

Using the definition of the joint entropy and rearranging terms gives

$$\begin{aligned}RHS=&H(X,Y)\\&-\sum_{i=1}^{N_X}\log_2 P_X(i)\sum_{j=1}^{N_Y}P_{XY}(i,j)\end{aligned}$$

$$-\sum_{j=1}^{N_Y} \log_2 P_Y(j) \sum_{i=1}^{N_X} P_{XY}(i,j)$$

It was previously argued that

$$P_X(i) = \sum_{j=1}^{N_Y} P_{XY}(i,j)$$

$$P_Y(j) = \sum_{i=1}^{N_X} P_{XY}(i,j)$$

Therefore

$$RHS = -H(X,Y) - \sum_{i=1}^{N_X} P_X(i) \log_2 P_X(i) - \sum_{j=1}^{N_Y} P_Y(j) \log_2 P_Y(j)$$

$$RHS=H(X)+H(Y)-H(X,Y)=I(X,Y)$$

Lemma 6. If X and Y are statistically independent, then I(X,Y)=0.

Demonstration:

It has previously been argued that $P_{XY}(i,j) = P_X(i)P_Y(j)$ if X and Y are statistically independent. From Lemma 5, we have

$$I(X,Y) = \sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} P_{XY}(i,j) \log_2 \left\{ \frac{P_{XY}(i,j)}{P_X(i)P_Y(j)} \right\}$$

If X and Y are independent, the argument of $\log_2$ is identically one, and I(X,Y)=0.

Appendix 2. Jointly Gaussian data sets and the mutual information of jointly Gaussian data set pairs

We construct here a procedure for generating jointly Gaussian data sets $\{Y^1\}$ and $\{Y^2\}$ from two independent Gaussian data sets $\{X^1\}$ and $\{X^2\}$. This is followed by a demonstration showing that the mutual information of two jointly Gaussian data sets with a cross-correlation coefficient r is $-0.5\log(1-r^2)$.

For simplicity of presentation we consider the special case of data sets that have zero mean and equal variance. The procedure can be extended to the more general case. Data sets with these limitations are, however, sufficient if the investigation is limited to comparison tests of mutual information algorithms. Let $\{X^1\}=(x_1^1,x_2^1,x_3^1,\cdots\cdots x_N^1)$ and $\{X^2\}=(x_1^2,x_2^2,x_3^2,\cdots\cdots x_N^2)$ be Gaussian distributed with zero mean and the same variance σ^2. It is further assumed that they are uncorrelated, that is, their cross-correlation coefficient r is equal to zero. Given the assumption of zero correlation, their joint probability distribution is the product of their individual probability distributions.

$$P_{X^1X^2}(x^1,x^2)=\frac{1}{2\pi\sigma^2}\exp\left\{-\left[(x^1)^2+(x^2)^2\right]/2\sigma^2\right\}=\frac{1}{2\pi|\Sigma_X|^{1/2}}\exp\left\{-\underline{x}^T\sum\nolimits_X^{-1}\underline{x}/2\right\}$$

where Σ_X is the (X^1,X^2) covariance matrix.

$$\sum\nolimits_X=\begin{pmatrix}\sigma^2 & 0\\ 0 & \sigma^2\end{pmatrix}$$

Two data sets $\{Y^1\}=(y_1^1,y_2^1,y_3^1,\cdots\cdots y_N^1)$ and $\{Y^2\}=(y_1^2,y_2^2,y_3^2,\cdots\cdots y_N^2)$ with zero means, equal variance σ^2 and cross-correlation r are jointly Gaussian if their joint probability density function is

$$P_{Y^1Y^2}(y^1,y^2)=\frac{1}{2\pi|\Sigma_y|^{1/2}}\exp\left\{-\underline{y}^T\Sigma_y^{-1}\underline{y}/2\right\}$$

Σ_y is the (Y^1,Y^2) covariance matrix.

$$\sum\nolimits_Y=\sigma^2\begin{pmatrix}1 & r\\ r & 1\end{pmatrix}\qquad\qquad \sum\nolimits_Y^{-1}=\frac{1}{(1-r^2)\sigma^2}\begin{pmatrix}1 & -r\\ -r & 1\end{pmatrix}$$

Matrix A is a two-dimensional linear transformation relating $\{X^1\}$ and $\{X^2\}$, independent Gaussian random variables, to $\{Y^1\}$ and $\{Y^2\}$, jointly distributed Gaussian variables.

$$\begin{pmatrix} x_j^1 \\ x_j^2 \end{pmatrix} = A \begin{pmatrix} y_j^1 \\ y_j^2 \end{pmatrix}$$

By construction

$$-\underline{x}^T \sum\nolimits_X^{-1} \underline{x} = -\underline{y}^T \sum\nolimits_Y^{-1} \underline{y}$$

Using the expression for $\sum_X^{-1}$ and re-expressing $\underline{x}$ in terms of $A\underline{y}$ gives

$$\underline{x}^T \underline{x} = \underline{y}^T A^T A \underline{y} = \sigma^2 \underline{y}^T \sum\nolimits_Y^{-1} \underline{y}$$

Let A be given by

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

Using this representation for A and the expression for $\sum_Y^{-1}$ above gives

$$A^T A = \begin{pmatrix} a^2 + c^2 & ab + cd \\ ab + cd & b^2 + d^2 \end{pmatrix} = \sigma^2 \sum\nolimits_Y^{-1} = \frac{1}{1-r^2} \begin{pmatrix} 1 & -r \\ -r & 1 \end{pmatrix}$$

Solving for b, c and d in terms of a and r gives

$$A = \begin{pmatrix} a & -ar + \sqrt{1 - a^2(1-r^2)} \\ \sqrt{\dfrac{1-a^2(1-r^2)}{(1-r^2)}} & -r\sqrt{\dfrac{1-a^2(1-r^2)}{(1-r^2)}} - a\sqrt{(1-r^2)} \end{pmatrix}$$

There are an infinity of A's that depend on the choice of a. We use here the simplest case, a=1.

$$A = \begin{pmatrix} 1 & 0 \\ \dfrac{r}{\sqrt{1-r^2}} & \dfrac{-1}{\sqrt{1-r^2}} \end{pmatrix} \qquad A^{-1} = \begin{pmatrix} 1 & 0 \\ r & -\sqrt{1-r^2} \end{pmatrix}$$

In the next step, we need to establish the relationship cited in the text between mutual information $I(Y^1, Y^2)$ and r, the cross-correlation coefficient. In this derivation, we use the property that $\{Y^1\}$ and $\{Y^2\}$ are jointly distributed, have correlation r, and are related to independent Gaussian data sets $\{X^1\}$ and $\{X^2\}$ by linear transformation A. The derivation begins with the integral representation for mutual information expressed in terms of the joint and individual probability density functions. The integrals are taken from $-\infty$ to $+\infty$.

$$I(Y^1, Y^2) = \iint P_{Y^1Y^2}(y^1, y^2) \log\left\{ \frac{P_{Y^1Y^2}(y^1, y^2)}{P_{Y^1}(y^1) P_{Y^2}(y^2)} \right\} dy^1 dy^2$$

By construction, Y^1 and Y^2 are jointly Gaussian. Therefore,

$$P_{Y^1Y^2}(y^1, y^2) = \frac{1}{2\pi |\Sigma_Y|^{1/2}} e^{-\underline{y}^T \Sigma_Y^{-1} \underline{y}/2}$$

Σ_Y is the Y^1, Y^2 covariance matrix where as before equal variances are assumed.

$$\sum_Y = \sigma^2 \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix} \qquad \left|\sum_Y\right|^{1/2} = \sigma^2 (1 - r^2)^{1/2}$$

Y^1 and Y^2 are Gaussian distributed.

$$P_{Y^1}(y^1) = \frac{e^{-(y^1)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} \qquad P_{Y^2}(y^2) = \frac{e^{-(y^2)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}$$

This gives the following expression for mutual information

$$I(Y^1, Y^2) = \iint \frac{e^{-\underline{y}^T \Sigma_Y^{-1} \underline{y}/2}}{2\pi |\Sigma_Y|^{1/2}} \log\left\{ \frac{\dfrac{e^{-\underline{y}^T \Sigma_Y^{-1} \underline{y}/2}}{2\pi |\Sigma_Y|^{1/2}}}{\dfrac{e^{-(y^1)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} \dfrac{e^{-(y^2)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}} \right\} dy^1 dy^2$$

Given the previously stated expression for $|\Sigma_Y|^{1/2}$, the expression for mutual information simplifies to

$$I(Y^1, Y^2) = \iint \frac{e^{-\underline{y}^T \Sigma_Y^{-1} \underline{y}/2}}{2\pi\sigma^2 (1 - r^2)^{1/2}} \log\left\{ \frac{e^{-\underline{y}^T \Sigma_Y^{-1} \underline{y}/2}}{e^{-(y^1)^2/2\sigma^2} e^{-(y^2)^2/2\sigma^2} (1 - r^2)^{1/2}} \right\} dy^1 dy^2$$

This can be re-expressed as an integral in x^1 and x^2.

$$\underline{y} = A^{-1} \underline{x}$$

$$\begin{pmatrix} y^1 \\ y^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ r & -\sqrt{1 - r^2} \end{pmatrix} \begin{pmatrix} x^1 \\ x^2 \end{pmatrix} = \begin{pmatrix} x^1 \\ rx^1 - \sqrt{1 - r^2}\, x^2 \end{pmatrix}$$

$$dy^1 dy^2 = \left| \frac{\partial(y^1, y^2)}{\partial(x^1, x^2)} \right| dx^1 dx^2$$

where the right hand side of the last equation is the absolute value of the determinant of the Jacobian of the transformation.

$$dy^1dy^2 = \begin{vmatrix} 1 & 0 \\ r & -\sqrt{1-r^2} \end{vmatrix} dx^1dx^2 = \sqrt{1-r^2}\,dx^1dx^2$$

By construction $\underline{y}^T \sum_Y^{-1} \underline{y} = \underline{x}^T x/\sigma^2$. This gives

$$I(Y^1,Y^2) = \iint \frac{e^{-\underline{x}^T\underline{x}/2\sigma^2}}{2\pi\sigma^2} \log\left\{ \frac{e^{-\underline{x}^T\underline{x}/2\sigma^2}}{e^{-(y^1)^2/2\sigma^2}\, e^{-(y^2)^2/2\sigma^2}\, (1-r^2)^{1/2}} \right\} dx^1dx^2$$

Taking logarithms of the exponentials gives

$$I(Y^1,Y^2) = \iint \frac{e^{-\underline{x}^T\underline{x}/2\sigma^2}}{2\pi\sigma^2} \left\{ \frac{1}{2\sigma^2}\left(-\underline{x}^T\underline{x} + (y^1)^2 + (y^2)^2\right) - \log\sqrt{1-r^2} \right\} dx^1dx^2$$

An expression for $(y^1)^2 + (y^2)^2$ is constructed from the defining linear relationship between x's and y's.

$$\underline{y} = A^{-1}\underline{x}$$

$$\begin{pmatrix} y^1 \\ y^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ r & -\sqrt{1-r^2} \end{pmatrix} \begin{pmatrix} x^1 \\ x^2 \end{pmatrix} = \begin{pmatrix} x^1 \\ rx^1 - \sqrt{1-r^2}\,x^2 \end{pmatrix}$$

$$(y^1)^2 + (y^2)^2 = (x^1)^2 + \{rx^1 - \sqrt{1-r^2}\,x^2\}^2 = (x^1)^2 + r^2(x^1)^2 - 2r\sqrt{1-r^2}\,x^1x^2 + (1-r^2)(x^2)^2$$

This expression in the integral becomes

$$-\underline{x}^T x + (y^1)^2 + (y^2)^2 = -(x^1)^2 - (x^2)^2 + (x^1)^2 + r^2(x^1)^2 - 2r\sqrt{1-r^2}\,x^1x^2 + (1-r^2)(x^2)^2$$

$$= r^2(x^1)^2 - r^2(x^2)^2 - 2r\sqrt{1-r^2}\,x^1x^2$$

The integral for mutual information becomes

$$I(Y^1,Y^2) = \iint \frac{e^{-\underline{x}^T\underline{x}/2\sigma^2}}{2\pi\sigma^2} \left\{ \frac{1}{2\sigma^2}\left(r^2(x^1)^2 - r^2(x^2)^2 - 2r\sqrt{1-r^2}\,x^1x^2 \right) - \log\sqrt{1-r^2} \right\} dx^1dx^2$$

Consider the integral

$$\iint \frac{e^{-\underline{x}^T\underline{x}/2\sigma^2}}{2\pi\sigma^2} \left\{ \frac{1}{2\sigma^2}\left(r^2(x^1)^2 - r^2(x^2)^2 \right) \right\} dx^1dx^2$$

The two terms are of equal magnitude and opposite sign, and the double integral is therefore equal to zero. Similarly consider

$$\iint \frac{e^{-\underline{x}^T\underline{x}/2\sigma^2}}{2\pi\sigma^2} \left\{ \frac{1}{2\sigma^2}\left(-2r\sqrt{1-r^2}\,x^1x^2 \right) \right\} dx^1dx^2$$

Each integral is of an odd function over the range $-\infty$ to $+\infty$ and is therefore equal to zero. The integral for mutual information simplifies to

$$I(Y^1, Y^2) = -\iint \frac{e^{-\underline{x}^T \underline{x}/2\sigma^2}}{2\pi\sigma^2} \left\{ \log\sqrt{1-r^2} \right\} dx^1 dx^2$$

Using

$$\int_{-\infty}^{+\infty} e^{-z^2/2\sigma^2} = (2\pi)^{1/2}\sigma$$

gives

$$I(Y^1, Y^2) = -\log\sqrt{1-r^2} = -\frac{1}{2}\log(1-r^2)$$

Appendix 3. Binary representation of XY partitioning and generalization to embedded data

The previous section provided details of the local adaptive partitioning used by Fraser and Swinney to calculate mutual information. The space being partitioned is that of the joint distribution of $X = \{x_1, x_2, \cdots x_N\}$ and $Y = \{y_1, y_2, \cdots y_N\}$, a subset of the XY plane which may be considered a two-dimensional embedding space whose elements are (x_i, y_i), i=1,2,....N. The following steps are used to implement the procedure:

1. Let the number of elements of both X and Y be $N = 2^n$ (the binary logic of the algorithm requires $N = 2^n$).
2. Rank order both X and Y with no repeated elements so that they both map to permutations of the integers 0, 1, ..., 2^n-1. To avoid repeated elements, one may assign higher ranks to numbers appearing earlier in the series. Call these rank-ordered lists $X^R = \{x_1^R, x_2^R, \cdots x_N^R\}$ and $Y^R = \{y_1^R, y_2^R \cdots y_N^R\}$. X^R and Y^R are equiprobable.
3. Transform the elements of X^R to binary. Since the $0 \leq x_k^R \leq 2^n - 1$, these binary representations have at most n bits – i.e., $x_k^R = a_k^{n-1} a_k^{n-2} \cdots a_k^0$. Here, a_k^{n-1} is the most significant bit of x_k^R, a_k^{n-2} the second most significant, etc. Perform the same transformation on the elements of Y^R to get $y_k^R = b_k^{n-1} b_k^{n-2} \cdots b_k^0$.
4. Interleave the bits of x_k^R and y_k^R to get

$$z_k^R = (a^{n-1}{}_k b_k^{n-1} a_k^{n-2} b_k^{n-2} \cdots a_k^0 b_k^0) . \qquad (1)$$

The two leftmost elements of z_k^R are the most significant bits of x_k^R and y_k^R, respectively, the next two are the next most significant bits, etc. For example, suppose $(x_k^R, y_k^R) = (5,47)$. Then, using the binary representations, 5 = 000101 and 47=101111, the interleaved representation of (x_k^R, y_k^R) is

$$(x_k^R, y_k^R) \Rightarrow z_k^R = (010001110111) .$$

A crucial advantage of this representation derives from the observation that the successive bit pairs provide a tree representation for the location of (x_k^R, y_k^R) in the two-dimensional embedding space. To see this, label the axes of a two-dimensional embedding space by x and y and consider the region

$0 \le x, y \le 2^5 - 1$. If this region is subdivided into 4 quadrants as in Figure 13a, then the bottom-left quadrant contains all those vectors with six-bit x's whose most significant bits are 0 and with y's whose most significant bits are also zero, the bottom-right quadrant contains all those x's whose most significant bits are 1 and those y's whose most significant bits are 0, etc. The location of any interleaved point in this subdivision is thus labeled by its first two elements; the (x_k^R, y_k^R) in our example is in quadrant 01. If this quadrant is again subdivided into four, the next two bits of z_k^R specify its location in the new subdivision (Figure 13b), and so on.

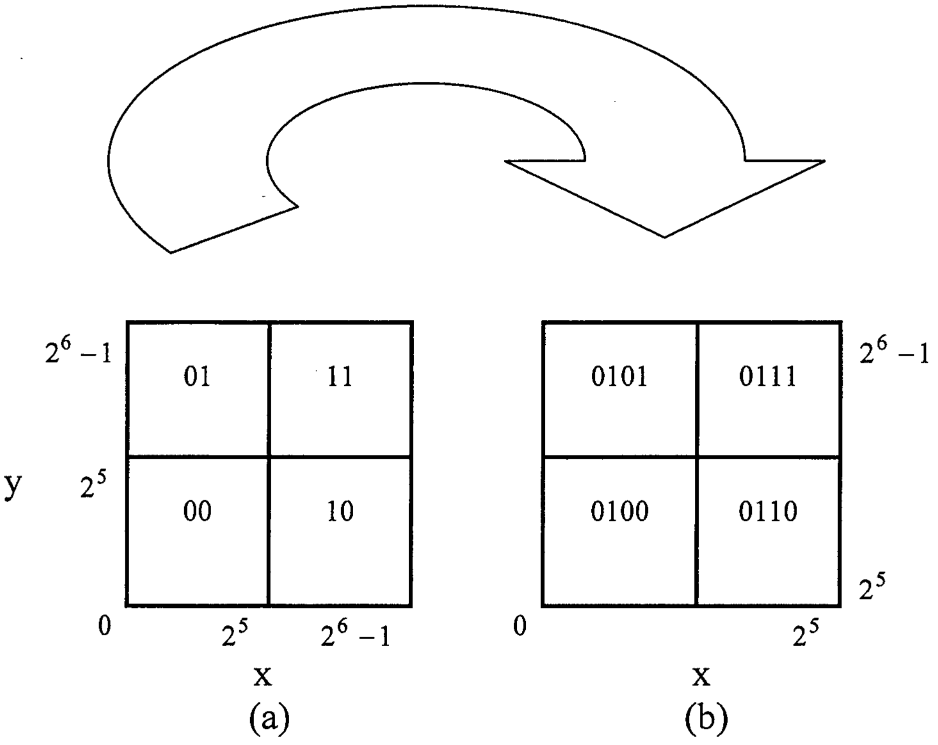

Figure 13. (a) Partition of $0 \le x,y \le 2^6\text{-}1$ into 4 quadrants. (b) Partition of quadrant 01 (upper left) into four sub-quadrants.

The technique of interleaving may also be used to implement time-delay embedding. Consider the m-dimensional embedding of X with a specified lag

$$X = (x_k, x_{k+Lag}, x_{k+2Lg}, \cdots\cdots x_{k+(m-1)Lag})$$

Using the notation of Equation. (1), the m-dimensional embedding vector, X_k, may be represented as

$$X_k \rightarrow u_k = (a_k^{n-1} a_{k+Lag}^{n-1} \cdots a_{k+(k-1)Lag}^{n-1})(a_k^{n-2} a_{k+Lag}^{n-2} \cdots a_{k+(k-1)Lag}^{n-2}) \cdots\cdots (a_k^{0} a_{k+Lag}^{0} \cdots a_{k+(m-1)Lag}^{0}) \qquad (2)$$

a number that uniquely represents X_k. A similar embedding and interleaving of Y gives $Y = (y_k, y_{k+Lag}, y_{k+2Lg}, \cdots\cdots y_{k+(m-1)Lag})$ and

$$Y_k \rightarrow v_k = (b_k^{n-1} b_{k+Lag}^{n-1} \cdots b_{k+(k-1)Lag}^{n-1})(b_k^{n-2} b_{k+Lag}^{n-2} \cdots b_{k+(k-1)Lag}^{n-2}) \cdots\cdots (b_k^{0} b_{k+Lag}^{0} \cdots b_{k+(m-1)Lag}^{0})$$

The interleaved sets, $\{u_k\}$ and $\{v_k\}$, each consists of 2^n numbers, each number specified by $n \times m$ bits. To calculate the mutual information of *X* and *Y* $\{u_k\}$ and $\{v_k\}$ are converted to decimal and used as inputs in the Fraser-Swinney algorithm.